# Driver Exposure to Particulate Matter: Field Study, Data Analysis and Modelling

Birgit Krausse

**A thesis submitted in partial fulfilment of the requirements of the De Montfort University for the degree of Doctor of Philosophy**

***July, 2004***

Institute of Energy and Sustainable Development
De Montfort University Leicester

# Driver Exposure to Particulate Matter: Field Study, Data Analysis and Modelling

Birgit Krausse

*Gesell` dich einem Bessern zu,*
*dass mit ihm deine besser`n Kräfte ringen.*
*Wer selbst nicht weiter ist als du,*
*der kann dich auch nicht weiterbringen.*

FRIEDRICH RÜCKERT (1788 - 1866)

I dedicate this thesis to the memory of my father who encouraged me to seek new challenges and always believed that I would achieve my goals.

# Abstract

An empirical study investigating human exposure to particulate matter in a transport microenvironment is described in this thesis. The focus of the investigation is the exposure of car drivers to ultrafine particles on urban roads. In-vehicle exposure to larger particles ($PM_{10}$, $PM_{2.5}$) is also explored.

The field study methodology, developed particularly for this project and applied during a twelve month field campaign, is outlined in some detail. The data processing methods and the structure of the field study database are described.

Findings from the analysis of high-frequency time series data are presented. Results indicate there are significant differences between ultrafine particles and particles of the larger size ranges ($PM_{10}$, $PM_{2.5}$) in terms of their short-term variability and sensitivity to external parameters (i.e. meteorological conditions and traffic events). Further analysis reveals significant differences in seasonal and diurnal variability of average exposure values for the investigated particle size ranges. Novel visualisation techniques are described which aid the detection of significant patterns in large volumes of time series data.

The methods and findings regarding the analysis of the main determinants of driver exposure by multiple linear regression analysis are described. Exposure levels are shown to be strongly affected by particular road layouts and meteorological conditions. Different determinants are identified for ultrafine particles than for the other two size ranges.

Three scenarios are presented which aim to utilize field data in conjunction with the findings regarding the main determinants to estimate exposure on future journeys. Results show that strong over- or underestimation of exposure levels occurs in some cases, indicating that alternative modelling methods may be more suitable.

Finally, a method for modelling average annual driver exposure is proposed. Methods and implications are discussed. The results indicate that the method may produce useful results but validation with additional data is required.

# Acknowledgements

I would like to express my gratitude to the De Montfort University Leicester, and particularly to the Institute of Energy and Sustainable Development (IESD), who provided me with the opportunity and funding to carry out this research. I also gratefully acknowledge the contribution of data by the Pollution Control Group of Leicester City Council and the equipment funding provided by the EPSRC.

Special thanks are due to my supervisor John Mardaljevic for his advice and guidance throughout the project, and my colleague and friend Kevin Turpin, with whom I shared the long days of the field campaign. I would also like to thank the staff at the IESD for their continuous support and assistance.

I am especially grateful for the encouragement given to me by my family over the years. Without their patience and support I would not have been in a position to carry out this research.

Finally, I thank Brian for always being there for me and for providing excellent advice and software support.

**I declare that the content of the submission represents solely my own work.**

**Birgit Krausse, July 2004.**

# Contents

# Figures

## Chapter 1. *Introduction*

## Chapter 2. *Particulate Matter and Personal Exposure: Legislation and Scientific Background*



## Chapter 3. *Field Campaign: Data Collection and Processing*



## Chapter 4. *Overview and General Data Analysis*

## Chapter 5. *Identifying the Main Determinants*



## Chapter 6. *Predictive Modelling*



## Chapter 7. *Discussion and Conclusions*

# Tables

## Chapter 1. *Introduction*

## Chapter 2. *Particulate Matter and Personal Exposure: Legislation and Scientific Background*



## Chapter 3. *Field Campaign: Data Collection and Processing*



## Chapter 4. *Overview and General Data Analysis*

## Chapter 5. *Identifying the Main Determinants*



## Chapter 6. *Predictive Modelling*



## Chapter 7. *Discussion and Conclusions*

## *Bibliography*

## *Glossary*

## Appendix A. *Auxiliary Material for Chapter 3*

## Appendix B. *Auxiliary Material for Chapter 4*



## Appendix C. *Auxiliary Material for Chapter 5*

## Appendix D. *Auxiliary Material for Chapter 6*

Chapter

# 1 Introduction

Personal exposure to air pollutants is widely recognised as an important issue in regards to human health in urban areas. One of the major sources contributing to urban air pollution is the transport sector. Many of the pollutants produced by the combustion engines of vehicles have been identified as presenting significant health hazards. Due to the introduction of catalytic converters, emissions of gaseous pollutants from cars have been reduced. More recently however, the focus of exposure studies has shifted to include particulate matter. In the UK, transport sources contribute about 25 % to the overall emissions of particulate matter and an even larger proportion in urban areas, i.e. 78 % in London (DETR, 2000a).

Epidemiological studies indicate that exposure to particulate matter can lead to adverse effects on human health. Although a clear causality has not yet been established, it is widely suspected that elevated concentrations of small particles, which can penetrate deep into the lungs, are responsible for negative health outcomes such as increased mortality, respiratory diseases as well as cardiovascular disease. Particularly vulnerable groups of the population such as the elderly or chronic sick are believed to be more at risk (e.g. COMEAP, 1995 and 2001; Pope III, 2000; Weijers et al., 2001). The introduction of air quality standards for particles has lead to an increased awareness regarding these issues. However, it has been suggested that the currently measured $PM_{10}$ and even $PM_{2.5}$ concentrations[1] do not represent the most dangerous particle fractions (i.e. greatest health risk). For example, Weijers et al.

1. Defined as particles which pass through a size selective inlet with a 50 % efficiency cut-off at 10 μm or 2.5 μm aerodynamic diameter, respectively (QUARG, 1996).

(2001) and Pope III (2000) recommend to focus on smaller size fractions and number concentration measurements (i.e. ultrafine particles, with a diameter < 0.1 μm) in order to take account of the potentially more damaging particles.

This distinction between particles of different size ranges is especially important in regards to traffic related particle emissions, where differences in spatial variability are observed for different sized particles. For example, Harrison et al. (1999a) found that $PM_{10}$ mass measured a short distance from the road is more representative of the local background concentration, while counts of ultrafine particles are highly sensitive to local traffic conditions. Findings published by Molnár et al. (2002) further support this. They report that none of the particle mass fractions, measured at a distance of 6 m from the road side, correlated with ultrafine number concentrations, and thus suggest that neither $PM_{10}$ nor $PM_{2.5}$ concentrations can be used to infer levels of ultrafine particles. Junker et al. (2000) found that the majority of particles < 0.421 μm measured at three locations in Basel, Switzerland, were of the ultrafine size fraction (on average 82 - 87 %), and that peak number concentrations were measured during periods of dense traffic in the early morning and late afternoon, i.e. rush hour periods.

Exposure research frequently uses the concept of 'microenvironments' to investigate the levels of pollutants humans are exposed to. Microenvironments are defined as zones that are believed to have distinct characteristics with respect to pollutant concentrations and/or variability (Sexton and Ryan, 1988). The commuter vehicle represents one urban microenvironment in which humans are potentially exposed to high pollutant concentrations on a regular basis. A number of studies comparing outdoor and in-vehicle concentrations have shown that particle mass concentrations inside a vehicle often exceed outside concentrations (Gee et al., 1999; Adams et al., 2001a; Kingham et al., 1998).

A number of studies have attempted to quantify exposure to traffic related pollutants and also to relate the exposure to the travel mode (e.g. Adams et al., 2001a; Gómez-Perales et al., 2004; Gulliver and Briggs, 2004). Adams et al. (2001a) found that cyclists experienced lower $PM_{2.5}$ exposure levels than bus or car passengers while Gulliver and Briggs (2004) report higher exposure values for car drivers than for pedestrians for $PM_{10}$ but not for $PM_{2.5}$ or $PM_1$.

Some exposure studies, often focussed on an individual transport microenvironment, were carried out on pre-defined routes, which allowed an investigation of the variability of personal exposure depending on parameters such as vehicle speed, route travelled and meteorological conditions etc. (e.g. Dickens, 2000, EHHI, 2002, Alm et al., 1999, Adams et al., 2001b). For example, Alm et al. (1999) found that in-vehicle concentrations of fine particles (0.3 - 1.0 μm ) were affected by meteorological parameters as well as journey specific parameters, with wind speed and average vehicle speed being negatively correlated to concentration values. Opposite effects were observed for coarse particles (1 - 5 μm ). Adams et al. (2001b) identified route and wind speed as significant determinants of driver exposure to $PM_{2.5}$.

Most of the commuter studies noted above used integrated pollution measurements, i.e. typically by estimating hourly averages from total concentration or exposure values measured over a number of journeys. However, it has been shown by Dickens (2000) that the exposure to pollutants does not only depend on travel mode but also varies significantly for different types of road layout and location (i.e. motorway, urban road, etc.) and particularly vehicle speed. A typical commuter study therefore gives an indication of the exposure of individuals on particular journeys, but its findings may only be of limited use in estimating exposure for other journeys. Furthermore, most previous studies on exposure to particulate matter have focussed on the larger particle size fractions (i.e. $PM_{10}$, $PM_{2.5}$).

Taking account of these considerations, the present study was designed to record instantaneous particle exposure levels experienced by car commuters. During an extensive field study campaign, high frequency personal exposure data were collected in a private car while it was driven along three pre-defined urban routes. The routes were devised so that they included different road types, e.g. dual carriage way, residential streets, one way system with bus lane. The routes were split into individual road links to represent parts of the route with constant traffic flow and road conditions.

The main aim of the study was to investigate driver exposure to ultrafine particles in terms of its temporal dynamics and possible relationships with external parameters such as road type and meteorological conditions. Concentrations of $PM_{10}$ and $PM_{2.5}$ particles were also measured and analysed in order to reveal differences and similarities in exposure between the particles from different size ranges.

It was envisaged that the high resolution of the field data would be utilized to study in-vehicle exposure patterns in relation to transient traffic events as well as seasonal and diurnal variations. The study also intended to investigate the potential of the field study data for exposure modelling.

Novel visualisation techniques were explored. They aid the inspection of multiple time series data sets and thus enable the detection of significant patterns in large volumes of data. A set of custom written programming routines was developed to support the processing and analysis of the field data set.

Chapter 2 gives an overview of the most important issues related to particulate matter air pollution and personal exposure research. The development and execution of the field campaign are outlined in Chapter 3, together with a description of the programming methods used for data processing and analysis. The first part of the data analysis, which focusses on the dynamics of driver exposure, is the subject of Chapter 4. Results from a number of analysis methods are presented, including descriptive statistics for the field data set and time-series analyses which utilize novel visualisation techniques. Chapter 5 describes the methods and results of the multiple linear regression analysis which was used to identify the main determinants of driver exposure to particulate matter. The findings are then applied in Chapter 6 in an attempt to model road link and journey exposure. This chapter also describes a proposed method for estimating annual exposure using the field study data. Chapter 7 contains a discussion of the results and the main conclusions together with suggestions for further work. A glossary is included which provides short descriptions of frequently used terms and abbreviations. The appendices hold auxiliary information on the methods applied in the individual chapters, where necessary, and additional charts and tabulated results.

Chapter

# 2

# *Particulate Matter and Personal Exposure: Legislation and Scientific Background*

It is widely recognised that road transport represents one of the main sources of air pollution in the UK (APEG, 1999). Many pollutants which are produced by the combustion engines of vehicles have been identified as being major health hazards. Research has therefore increasingly focussed on pollutant concentrations in ambient air, particularly in urban areas where population density and hence exposure of humans to harmful pollutants is highest. One of the major air pollutants emitted from traffic sources is particulate matter (PM), which has been shown to have adverse effects on human health (Weijers et al., 2001).

Due to the health implications associated with exposure to particulate matter, which are discussed in more detail below, national and international standards have been established in order to monitor and reduce emissions, particularly in urban areas. The current standards are based on mass concentrations of particles with an aerodynamic diameter up to 10 μm , which are referred to as $PM_{10}$[1] (QUARG, 1996). As outlined in the following sections, recent research findings indicate that using

1. More strictly, particles which pass through a size selective inlet with a 50 % efficiency cut-off at 10 μm aerodynamic diameter (QUARG, 1996).

measurements of the smaller size fractions, i.e. particles with a diameter up to 2.5 μm ($PM_{2.5}$) or smaller, and number concentrations, rather than mass concentrations, might be more advisable.

Personal exposure research aims to investigate the exposure of humans to air pollutants as experienced throughout the day and during various activities. It is generally accepted that emissions from traffic sources contribute significantly to overall exposure levels, particularly for individuals who spend long periods in dense traffic. A number of studies have shown that commuters are exposed to especially high concentrations of air pollutants, with particle concentrations playing an important role (e.g. Colvile et al., 2001).

Particle pollution and exposure research comprises many specialised fields. This chapter aims to set the scene by introducing the main issues. It is not intended to be an exhaustive literature review but rather provides general background information relevant to particle pollution and driver exposure. It contains selected references to numerous research and legislation documents.

The following sections outline the policies and regulations introduced in the UK to combat particle pollution, provide an overview of the sources and health effects of particulate matter and briefly describe the monitoring instrumentation available. This is then followed by a discussion of personal exposure research relevant to driver exposure.

## 2.1 Policy and legislation

Air pollution is caused by a wide range of anthropogenic activities, with industry and transport as the major sources. In order to improve air quality, emissions from all these sources have to be reduced as much as possible. To tackle this problem, the UK government has taken various legislative steps.

### 2.1.1 The National Air Quality Strategy and current standards

As part of the Environment Act 1995, the National Air Quality Strategy was developed which introduced UK wide air quality objectives and set health-based standards for the main air pollutants (Benzene, 1,3-butadiene, carbon monoxide, lead, nitrogen dioxide, sulphur dioxide and particulate matter). A revised version was published in January 2000 (DETR, 2000a) and has since been updated by an addendum (DEFRA,

2003a) in order to take account of the latest research findings regarding risk and health effects caused by the major pollutants.

These standards and objectives are subject to regular review to take account of the latest information on the health effects of air pollution and technical and policy developments. The standards are generally based on recommendations by the EPAQS (The Expert Panel on Air Quality Standards), which take into account the best available evidence of the effects of air pollution on human health. In some cases European Union limit values, usually derived from World Health Organisation guidelines (WHO, 2000), are used to define UK pollution standards. The objectives of the Air Quality Strategy state policy intentions for the medium term, taking account of the costs and benefits and the feasibility and practicability of moving towards the standards. They are based on information on air pollution from the national air pollution monitoring network and the results from sophisticated air quality models.

The current air quality objectives for particulate matter, based on concentrations of $PM_{10}$, are for all parts of the UK, except London and Scotland, an annual mean of 20 $\mu g\ m^{-3}$ and a 24-hour mean of 50 $\mu g\ m^{-3}$ not to be exceeded for more than seven times a year. Both objectives are to be achieved by 31 December 2010 (DEFRA, 2003a).

In a report published in 1999, the APEG (Airborne Particles Expert Group) estimated that emissions from mainland Europe contribute significantly to the annual average $PM_{10}$ concentrations in the UK (15 % of primary and 20 % of secondary particles). However, as outlined in DETR (2000a), the major problems are caused by emissions from local sources: of the 186 kilotonnes of $PM_{10}$ emitted in the UK in 1999, about 38 kilotonnes came from residential heating, about 31.5 kilotonnes from road transport, about 30 kilotonnes from industrial sources and about 19 kilotonnes from public power. Although road transport is not the main source of $PM_{10}$ at the UK level, it tends to be the most significant source in urban areas where traffic congestion typically occurs. Moreover, road transport was found to the main contributor to $PM_{2.5}$ concentrations (approximately 20 %) (DEFRA, 2003a).

### 2.1.2 Local air quality management

Technological developments will help reduce emissions from individual vehicles, but the reduction of overall emissions from transport sources has to be tackled at a local

level. This forms the basis of the local air quality management guidance[2] (LAQM), under which local authorities are required periodically to review and assess the current and future air quality in their area and work towards achieving the objectives and standards set by the Air Quality Strategy. When one or more objectives are unlikely to be met in a certain area, local authorities must declare it an AQMA (Air Quality Management Area) and draw up an action plan, setting out what will be done to meet the objectives within this AQMA. In order to identify problem areas and evaluate solutions, local authorities use a variety of tools and approaches, such as air quality monitoring, local emissions inventories and numerical modelling (e.g. air dispersion models and traffic models). Solutions can include road planning measures as well as improvement of public transport (DEFRA, 2003b).

In order to evaluate the progress towards sustainable development, the Government has highlighted a number of 'Indicators for Sustainable Development' (UK Government, 1999). The 'Headline Air Quality Indicator' of a monitoring site measures the number of days on which the concentration of any one of five pollutants which are recognised as the most important for causing short-term health effects (i.e. carbon monoxide, nitrogen dioxide, ozone, $PM_{10}$ and sulphur dioxide) were 'moderate' or 'high' according to the 'Air Pollution Information Service' bandings. These bandings are based on health effects caused by air pollutants, with 'moderate' denoting "mild effects, unlikely to require action, may be noticed amongst sensitive individuals". 'High' and 'very high' indicate worse conditions.[3]

As noted above, current UK air quality standards are based on the $PM_{10}$ size fraction. Although a number of studies have recommended to introduce standards for smaller particles, which are thought to more strongly affect health endpoints (discussed in Section 2.2.3), EPAQS has reaffirmed that the $PM_{10}$ standard is a reasonable metric for the assessment of air quality in terms of particle pollution. However, this conclusion partly results from a lack of evidence that would be needed to define such a standard (DETR, 2000a), which is a result of particulate matter being a very complex pollutant. The varying composition of particulate matter in general and the great temporal and spatial variability of smaller particles hampers the identification

2. Current general guidance is published in LAQM.G1(00) (DETR, 2000b) and transport related air quality issues are dealt with in LAQM.G3(00) (DETR, 2000c).

3. From http://www.airquality.co.uk/archive/standards.php#band, accessed May 2004.

of causal relationships between particle concentrations and health effects. These issues are explored further in the following sections.

## 2.2 Particulate matter

### 2.2.1 Origins, physical parameters and composition

Particulate matter in the atmosphere originates either from primary emissions (i.e. emitted directly from a source) or secondary emissions (i.e. formed in the atmosphere through chemical reactions). The Third Report of the Quality of Urban Air Review Group (QUARG, 1996) named the three predominant contributors to particulate matter as being secondary aerosols, primary vehicle exhaust emissions and suspended soil and road dusts. In the past, industrial processes have been the major sources of particle pollution in the UK. However, with the growth of vehicle numbers and traffic density, transport is now considered to be the main particle source, particularly in urban areas. (e.g. Colvile et al., 2001; APEG, 1999)

Physical characteristics of particulate matter

Because suspended particulate matter is a mixture of particles of different sizes and compositions, a number of physical parameters, such as mass, size, surface area and number, are used in the investigation of particle distributions and their effects. It is important to note that, although atmospheric particles are often not spherical, their size is usually described by an 'equivalent' diameter $D_p$, which is defined as "The diameter of a spherical particle which will give identical geometric, optical, electrical or aerodynamic behaviour to that of the particle (non-spherical) being examined".[4] For particulate matter pollutants an aerodynamical equivalent diameter is typically used.

The distribution curves in Figure 2-1 show that the largest number concentrations are observed for particles smaller than 0.1 μm (known as ultrafine particles), while only particles with diameters above 0.2 μm contribute substantially to mass concentrations. Particles with large surface areas, with respect to their diameter, are found in the size range between 0.02 μm and 1 μm (Weijers et al., 2001). Table 2-1

4. From IUPAC Compendium of Chemical Terminology, http://www.iupac.org/publications/compendium/index.html, accessed June 2004.

gives an overview of the particles size ranges and terminology commonly used in current particle research.

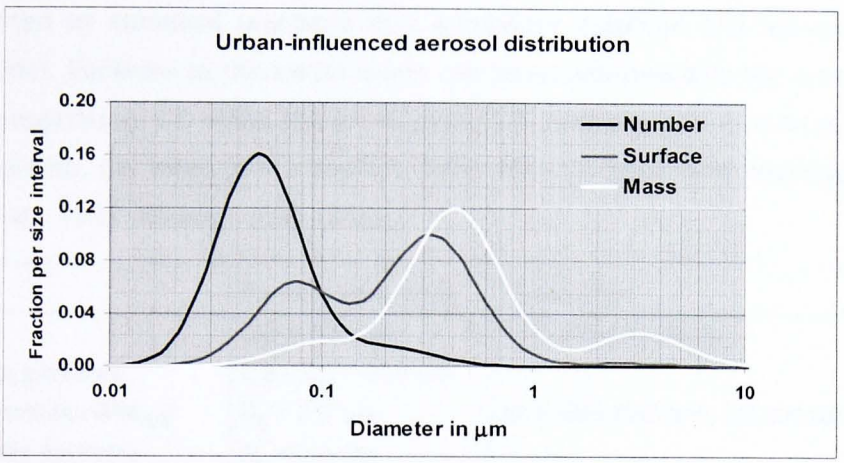


**Figure 2-1 Number, surface and mass distributions of urban air particles (From Weijers et al., 2001)**

The most commonly used metric for particle concentration measurements is $PM_{10}$, which is defined as particles which pass through a size selective inlet with a 50% efficiency cut-off at 10 µm aerodynamical diameter. $PM_{10}$ concentration measurements are therefore a good approximation of the mass of particles in the atmosphere which have a diameter smaller than 10 µm (APEG, 1999). However, according to Weijers et al. (2001), metrics based on other parameters, such as number concentrations, may be more appropriate since they better reflect biological mechanisms through which particle pollution is thought to cause health effects. In their review of other studies, Harrison et al. (1999) state that the median of the particle number concentrations measured in urban air clearly differs from the that of the mass concentration (0.09 - 0.12 µm and 0.4 - 0.8 µm, respectively), which indicates that mass concentration measurements such as $PM_{10}$ may not be a good indicator of ultrafine particle concentrations. Findings published by Molnár et al. (2002) support this further. They report that none of the measured particle mass fractions, not even the ones in the accumulation mode (approx. 0.1 - 0.4 nm), correlated with ultrafine number concentrations, and thus suggest that neither $PM_{10}$ nor $PM_{2.5}$ concentrations can be used to infer levels of ultrafine particles. The terms 'nuclei mode' and 'accumulation mode' refer to the chemical and mechanical processes by which particles are produced, i.e. nucleation from precursor gases, and coagulation & condensation, respectively.

When the very fine particles are first emitted they are in the nuclei mode but then rapidly combine to form larger particles which belong to the accumulation modes (see Figure 2-2). Some of the gaseous primary pollutants (i.e. sulphur and nitrogen oxides) are converted by chemical reactions into secondary sulphate and nitrate particles (nuclei mode). Particles in the nuclei mode can be transferred into the accumulation mode by coagulation, i.e. when two small particles combine to form a larger particle, or condensation, i.e. when gas or vapour molecules condense onto existing particles (Wilson et al., 1997; Wehner et al., 2002).

| Term | Particle size range | Comments |
|---|---|---|
| **$PM_{10}$** | $D_p$[a] < 10.0 μm | air quality standards |
| **coarse particles** | 2.5 < $D_p$ < 10.0 μm | |
| **fine particles ($PM_{2.5}$)** | $D_p$ < 2.5 μm | air quality standard, US but not UK |
| **ultrafine particles** | $D_p$ < 0.1 μm | |
| **nanoparticles** | $D_p$ < 0.05 μm | |
| **nuclei mode** | 0.005 < $D_p$ < 0.05 μm | typical size ranges for particles from vehicle emission (Kittelson, 1998) |
| **accumulation mode** | 0.1 < $D_p$ < 0.3 μm | |

**Table 2-1 Terminology for the most commonly used particle size ranges**

a. aerodynamic particle diameter

The size ranges given for the nuclei and accumulation modes in Table 2-1 are those used to describe emissions from vehicles, which dominate urban aerosol. However, when investigating particle emissions from individual sources, accumulation mode particles may be found in smaller size ranges. Generally, the accumulation mode contains most of the mass of sub-micrometer particles, while the nuclei mode contains most of the particle number (Kittelson et al., 1998).

The composition and physical parameters of suspended particulate matter vary greatly depending on the sources from which the particles are emitted. The size of a particle influences its environmental impact in various ways, such as atmospheric residence time, the ability to participate in atmospheric chemistry and the health effects it may cause.

Road transport produces particles of various sizes and chemical compositions. Particles generated by mechanical processes such as naturally occurring dust, resuspended road dust and products of tyre and brake wear tend to be in the coarser size fractions, i.e. > 2.5 μm, whereas primary combustion particles are generally

smaller, i.e. < 2.5 μm, and often well below 1 μm (APEG, 1999). Figure 2-2 shows the particle size distribution observed in close proximity to traffic sources, e.g. within the traffic flow.

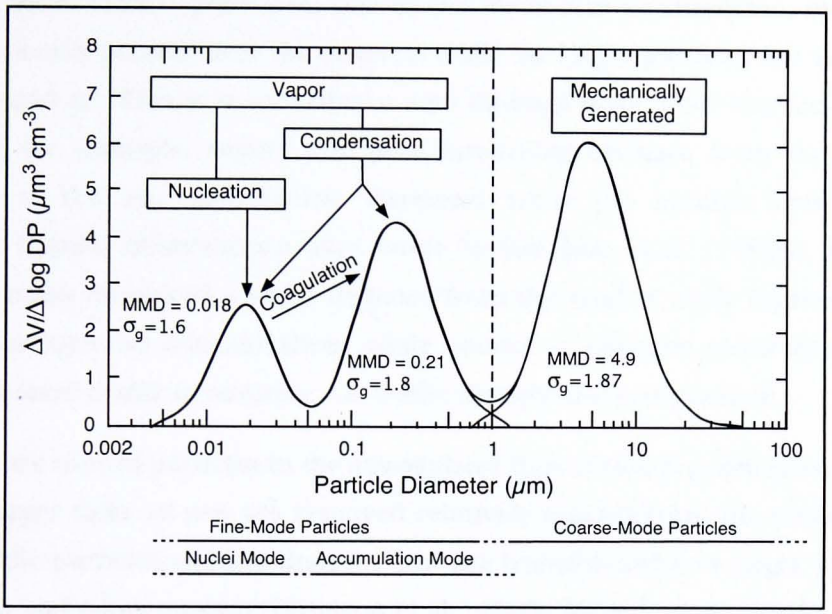


**Figure 2-2 Particulate matter size distribution collected in urban traffic flow showing formation mechanisms for particle modes (Re-drawn from Wilson et al., 1997)**

Studies investigating the size distribution of particle emissions have shown that most particles emitted by petrol and diesel vehicles are in the sub-micrometer range with petrol vehicles emitting even smaller particles than diesel vehicles (e.g. Weijers et al., 2001). Wehner and Wiedensohler (2003) found that particles in the range 0.01 - 0.1 μm were most strongly influenced by car traffic. Road side measurements carried out by Junker et al. (2000) found that ultrafine particles (< 0.1 μm) contributed on average 82 to 87 % to the total number counts of particles < 0.421 μm and reached maximum values during rush hour periods when they contributed up to 97 % to the total number counts. According to Hitchins et al. (2000), previous studies showed that the majority of exhaust particles from diesel engines were found in the range 0.02 - 0.13 μm, while particles from petrol engines were in the range 0.04 - 0.06 μm. Similar results were published by Charron and Harrison (2003) who reported that accumulation mode particles showed an association with heavy-duty traffic (mainly diesel vehicles) while particles in the range 0.03 - 0.06 μm could be linked to light-duty traffic (i.e. passenger cars). Kittelson et al. (2000) measured diesel emissions under real-world conditions and found that sub-

micrometer particles are bimodal, with a nuclei mode in the 10 - 20 nm range containing 60 to 95 % of the particle number counts and an accumulation mode (40 - 60 nm) containing 90 to 99 % of the particle mass.

As Wilson et al. (1997) point out, due to the formation mechanisms, nuclei mode particles are only present near the sources, while the larger particles are also present in background air. This is in accordance with findings from other studies. Wehner et al. (2002), for example, report that with increasing distance from the road, the maximum of the size distribution increased while the number concentrations decreased. Similar observations were made by Harrison et al. (1999a), who found that $PM_{10}$ mass measured a short distance from the road is more representative of the local background concentration, while counts of ultrafine particles are highly sensitive to local traffic conditions, i.e. traffic density and composition.

The residence time of particles in the atmosphere thus varies depending on their size. Particles larger than 10 µm are removed relatively quickly from the atmosphere by settling, while particles smaller than 0.1 µm are transformed into larger particles by coagulation and condensation (Kittelson et al., 1998; Wehner et al., 2002).

Influence of meteorological conditions on particle concentrations

Meteorological conditions have a strong influence on the particle concentrations and size distribution. Wehner and Wiedensohler (2003) report higher particle number concentrations in winter, compared to summer, which they attribute to reduced dispersion of emissions from traffic and heating sources, caused by atmospheric inversion layers and low wind speeds. Similar results were reported from other studies. For example, Charron and Harrison (2003) found that increased wind speeds cause a greater dispersion of $PM_{10}$ and particles with a diameter greater than 30 nm, leading to lower mass and number concentrations.

Monitoring studies have also shown that precipitation seems to favour high particle number concentrations of ultrafine particles and that high relative humidity levels are associated with increased particle sizes, due to hygroscopic growth and stronger coagulation effects (Bukowiecki et al., 2002; Harrison et al., 1999b; Wilson et al., 1997, USEPA, 2001). The mechanisms involved in formation and growth rates of ultrafine particles are very complex and have been the subject of a large number of studies in recent years. For example, Kulmala et al. (2004) concluded, based on a

review of numerous studies, that temperature and vapour concentrations strongly affect the formation and growth rate of ultrafine particles and that particle growth rates tend to be higher in summer than in winter. These processes are also thought to, at least partly, explain why particle number concentrations are typically highest in the morning. Charron and Harrison (2003) linked this phenomenon to a shift in the mode diameter of the distribution from 21 nm to 31 nm (morning and evening rush hour, respectively). Kittelson et al. (2000) found that increased concentrations of nanoparticles were measured at lower ambient temperatures, while Wehner and Wiedensohler (2003) linked peak number counts of particles in the range 10 - 15 nm to increased global radiation, which is thought to photochemically induce nucleation of gaseous components.

Particle mass concentrations ($PM_{10}$) have also been shown to vary depending on season (i.e. weather conditions). According to Harrison et al. (1997), in the winter months $PM_{10}$ consists mainly of $PM_{2.5}$ particles and can be shown to correlate with $NO_x$[5], which is used as a marker for traffic emissions. During the summer months, however, $PM_{10}$ consists mainly of coarse particles, which originate from resuspended road dust and soil. Those coarse particles were found to be positively correlated with wind speed, due to increased wind-driven resuspension particularly in the dryer months, while concentrations of fine particles decreased with increasing wind speeds. Similar seasonal variations in $PM_{10}$ and $PM_{2.5}$ concentrations were reported by Chan et al. (1997), who reported slightly increased concentrations of fine particles in winter and of coarse particles in summer, and Deacon et al. (1997) who mentions a potential link between increased summer $PM_{10}$ levels and photochemically generated secondary aerosols.

As noted in APEG (1999), $PM_{10}$ concentrations in the UK are strongly influenced by long-range transport of suspended particulate matter from mainland Europe. These particles contribute mainly to $PM_{10}$ and $PM_{2.5}$ background concentration. Concentrations of the smaller particles are less affected due to shorter residence times.

With traffic being one of the main sources of particle pollution, it is important to note that vehicle emissions can vary significantly depending on ambient temperature. A

5. $NO_x$ - oxides of nitrogen, often used as marker of traffic emissions, correlate well with small particles (e.g. Harrison et al., 2001; APEG, 1999; AEAT, 1998; Colvile et al., 2001)

report by AEAT (2001) provides quantitative results for the effect of ambient temperature on cold start emission levels. Measured emissions of ultrafine particles were approximately five times higher for a cold engine (i.e. cold start emissions at approximately -7 °C) than the corresponding hot start emissions (25 °C), for both petrol and diesel powered vehicles. Particle mass was even more strongly affected. Here, cold start emissions were shown to increase by more than an order of magnitude when the ambient temperatures were reduced from 20 °C to -7 °C. Although hot start emissions were relatively unaffected, cold start emissions are expected to contribute significantly to increased particle concentrations during periods with low ambient temperatures.

Chemical composition

The major components of particulate matter are sulphates, nitrates, carbonaceous compounds, water, hydrogen ions, ammonium ions (mainly fine particles) and materials originating from the earth's crust (coarse particles) (USEPA, 2001).

Particles emitted from road traffic consist predominantly of carbonaceous compounds and metals, with organic and elementary carbon making up more than half of the emitted mass. For example, a chemical analysis of $PM_{10}$ carried out by Harrison et al. (1997), found that 18 % of the particle mass could be attributed to elemental carbon and 20 % to organic carbon compounds. Other components of particle emissions from traffic include PAH (polycyclic aromatic hydrocarbons), sulphates and nitrates which are usually found in the size fraction < 0.5 μm (Weijers et al., 2001).

Particulate matter in traffic

Emission levels from vehicles are not only influenced by ambient temperature, but also strongly depend on the vehicle type (e.g. type of fuel used), speed and acceleration. In order to estimate vehicle emissions, primarily for the use in air quality modelling, emission factors are used. These factors provide information on the amount of pollutants emitted per distance or time travelled (e.g. g $km^{-1}$ or g $h^{-1}$ respectively). Traditionally, emission factors were based on average emission values for different vehicle types and driving patterns (i.e. driving behaviour categorised for different types of roads). More recently, research has focussed on drive cycles which are used to represent real-world driving conditions as explained in, for example, Ericsson et al. (2001), Gramotnev et al. (2003), Sturm et al. (1998). An example of an

emission factor matrix for $NO_x$, from the Modem project (Joumard et al., 1995), is shown in Figure 2-3. The matrix illustrates the effect of speed and acceleration on emission factors, with higher speeds and acceleration values being linked to increased emissions of $NO_x$.

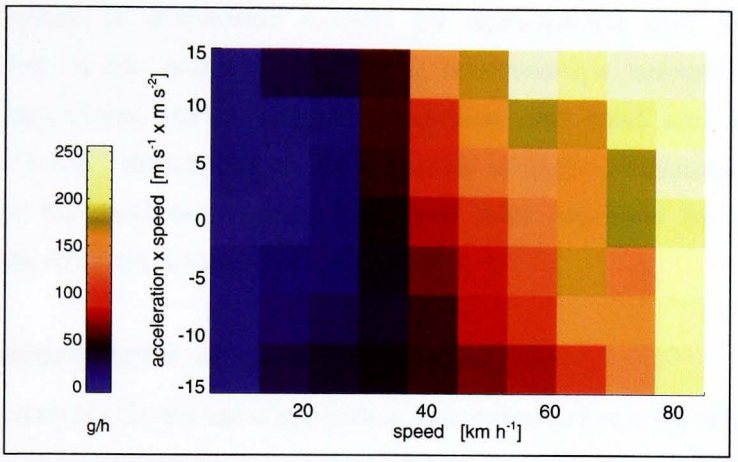


**Figure 2-3 Modem emissions matrix for gasoline vehicles[a] (vehicle class: passenger car, i.e. ECE 1503 <1.4 l)**

a. Data provided by TRL, Wokingham, UK.

Based on the strong correlation of $NO_x$ with traffic emissions of particles (e.g. Harrison et al., 1997), it can be assumed that vehicles emit higher concentrations of particles at higher speeds. However, due to the increased dilution at higher speeds, these emission characteristics are not always apparent when concentrations are measured. For example, Weijers et al. (2004) found that elevated particle concentrations were strongly dependent on traffic density, which is highest during periods of congestion, i.e. associated with low speeds, while Longley et al. (2003) report direct or exponential relationships of particle concentrations to traffic flow. Similarly, Junker et al. (2000) attribute differences in particle number concentrations between locations (in the same city) to differences in traffic density. Moreover, they link peak number concentrations to peak traffic densities in the early morning and late afternoon, i.e. rush hour periods.

A particular characteristic of air pollution from traffic sources is its 'source-receptor relationship'. In contrast to emissions from stationary emission sources, such as industrial point sources, road users (i.e. human receptors) are in very close proximity to the emission source. As noted above, number concentrations of fine and ultrafine

particles are closest near roads and decrease quickly with increasing distance from the road (e.g. Hitchins et al., 2000; Wehner et al., 2002; Harrison et al., 1999a). This means that road users are exposed to particularly high concentrations of particulate matter. As Colvile et al. (2001) point out, these concentrations can vary for different road users, due to the great variability of emissions from vehicles in the traffic stream and the influence of dispersion caused by atmospheric and vehicle induced turbulence. Due to the complex underlying relationships between traffic related particle concentrations, meteorological conditions and road and vehicle specific parameters, it is very difficult to model exposure to traffic emissions. Many studies have therefore focussed on investigating road user exposure by using personal exposure measurements, as outlined in Section 2.3.

### 2.2.2 Measurement of particulate matter

The type of measuring device used in particle concentration studies depends on which size fraction is to be investigated. Concentrations of larger particles are measured as mass concentrations while number concentrations are used for smaller particles.

Mass concentration values are usually obtained with gravimetric devices, such as the Partisol monitor, which measures the weight of particles on a filter. Direct-reading gravimetric devices providing continuous measurements, normally logged as 1 hour averages, include 'Tapered Element Oscillating Microbalance' (TEOM) systems and 'β-Attenuation Mass' monitors (BAM). The former use the principle of deposited particle mass changing the frequency of mechanical oscillation of a tapered glass tube while the latter measure the attenuation of beta-particles by airborne particles collected on a filter tape to give mass concentration values (APEG, 1999). However, there are also some optical instruments which can be used to measure mass concentrations, such as photometers and optical particle counters (OPC). The latter actually count particles and also obtain information on their size by measuring the intensity and/or angle of scattered light. Particle mass is then calculated using scaling factors (Keady, 2000; Turnkey Instruments, 2001). Very small particles cannot be detected with gravimetric devices (due to their low mass) or optical particle counters (due to their small size). Instead they are measured using condensation particle counters (CPC) which increase the size of a particle by condensing vapour onto it so that it 'grows' to sufficient size to be detected optically (TSI, 2000).

Due to the use of different techniques, the instruments have different detection limits and outputs. For a comparison of three optical devices see Keady (2000) and for a description of other particle detection instruments refer to Sarnat et al. (2003).

Most portable monitors used for personal exposure studies are integrated samplers that provide total concentration values for the sampling period from which average concentrations can be inferred. More recently optical devices have become available which allow the measurement of instantaneous particle concentration at short time intervals, often down to one second.

The TEOM is the instrument of choice for ambient air quality monitoring in the UK. Most local authorities use this device for continuous air quality monitoring. The Department of Transport, Environment and the Regions (DETR) advised that a correction factor be applied to the TEOM data in order to correct for the under-estimation caused by the loss of semi-volatile aerosols from its heated element (Ayers et al., 1999). This factor is now recommended as part of the standard TEOM monitoring procedure which is outlined in the technical guidance for local air quality management (DEFRA, 2003b). Other gravimetric instruments, such as the Partisol, provide more accurate results but require longer averaging periods. Both the TEOM and the Partisol are not suitable for personal exposure studies due to their size (i.e. not portable) and high instrument costs. Optical particle counters such as the OSIRIS from Turnkey Instruments Ltd. are increasingly used to measure concentrations $PM_{10}$ and $PM_{2.5}$ (AEAT, 2004). The main advantage of the OSIRIS is that it can be used for measurements with very short logging intervals (minimum of one second), which allows the investigation of short-term variability of particle concentrations, an important aspect of personal exposure research. However, as pointed out in DEFRA (2003b), such monitoring devices are based on light scattering techniques (i.e. are less accurate than gravimetric methods) and, although considered suitable for screening studies, they are not recommended for detailed assessments. However, co-location studies have shown that the values measured with the OSIRIS are in close agreement with the TEOM. For example, Gulliver and Briggs (2004) found a strong correlation ($r^2 = 0.83$) for 15 minute averages, with the OSIRIS exceeding the TEOM values by a factor of only 1.03.

Due to the particle detection method used, neither the OSIRIS nor the above mentioned gravimetric devices can be used for the measurement of ultrafine particles.

Many studies use scanning mobility particle sizers (SPMS) to measure number concentrations of particles in various size ranges (e.g. Longley et al., 2003; Junker et al., 2000; Bukowiecki et al., 2003, Hitchins et al., 2000). These instruments use differential mobility analysers (DMA) to determine the size of particles and condensation particle counters (CPC) to determine their number. Other instruments include aerosol spectrometers, as used for example by Weijers et al. (2004) and Tuch et al. (2003), as well as stand alone CPCs, as used by for example Keywood et al. (1999) and Harrison et al. (1999). The latter, however, only measure particle number concentrations without providing information regarding size distributions.

The majority of these instruments are rather large and fragile and thus not suitable for the use in personal exposure studies, where instruments have to be portable so that measurements can easily be carried out independently of location. In some cases, mobile measurements are performed using mobile laboratories, i.e. vehicles equipped with monitoring devices which can be moved to investigate particle pollution at hotspots, or measure particle concentrations while in traffic (e.g. Weijers et al., 2001, Weijers et al., 2004; Bukowiecki et al., 2003).

Only in the last few years have portable condensation particle counters with short logging intervals become available, so that personal exposure studies can now benefit from real-time measurements of ultrafine particle concentrations. Although very few studies have so far included such measurements (Section 2.3), one of the most often used instruments seems to be the P-Trak Ultrafine Particle Counter (model 8525, TSI Inc.) (Sarnat et al., 2003). Although it is less sensitive than other CPCs, the P-Trak is easily portable and can log data at short intervals (minimum of 1 second), and is thus ideally suited for personal exposure studies. This type of monitor has been mainly used for personal exposure assessment in the USA (e.g. Chan et al., 2004) but is now also used as part of the DAPPLE[6] project (Arnold et al., 2004).

### 2.2.3 Health effects

Particle emissions have long been suspected to have adverse effects on human health. This has been confirmed by medical research and has resulted in the introduction of national air quality standards for particles (see Section 2.1.1). The current standards

6. 'Dispersion of Air Pollution and Penetration into the Local Environment', project details on www.dapple.org.uk, accessed June 2004.

are based on the concentration of $PM_{10}$ particles, which are capable of entering the respiratory tract and reaching the deeper parts of the lungs (QUARG, 1996). More recent findings, as outlined below, suggest that smaller particles may have more severe effects.

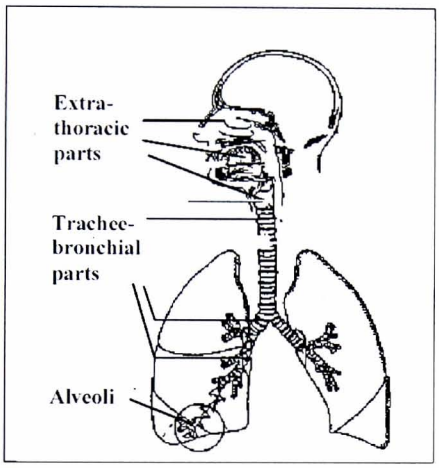


**Figure 2-4 The respiratory system (From Weijers et al., 2001)**

Coarse particles and part of the fine size fraction (0.5 - 2.5 μm ) are deposited in the extra-thoracic and the trachea-bronchial parts of the respiratory system (see Figure 2-4) and particles smaller than 1 μm can penetrate into the pulmonary alveoli. Based on the particle distribution shown in Figure 2-1, Weijers et al. (2001) conclude that most of the particle mass is deposited in the upper airways while the highest number deposition occurs in the alveoli. Since the number-size distribution reaches a maximum in the ultrafine range, this means that the particles getting deposited in the alveoli are small enough to be taken up in the vascular system, small enough to reach the interstitial spaces of the lung tissue (i.e. evading clearance by microphages) and have a small mass-median and a relatively large surface, such that toxins can be released quickly (Weijers et al., 2001). However, it is important to note that, despite evidence of clear statistical associations between particulate air pollution and health, the underlying biological mechanisms have not yet been clearly identified (e.g. COMEAP, 1995; Pope III, 2000; Weijers et al., 2001). The main problem is that suspended particles vary significantly in size, composition and origin. Moreover, it is suspected that adverse health effects are not solely caused by high mass or number concentrations of particles but also depend on the composition of the entire pollution

mixture, i.e. certain pollutants may enhance the detrimental health effects of particles (Weijers et al., 2001).

The three aspects that have to be considered when determining health effects of air pollutants are magnitude, duration and frequency of exposure. A large number of epidemiological studies have been carried out, particularly in the US, in order to investigate the effects of particulate air pollution on health. For example, Pope III (2000) gives an extensive summary on the state of epidemiology research regarding the health impacts of particulate matter. Epidemiological studies can be grouped into acute exposure studies, which evaluate the short-term variations in health end points (e.g. mortality counts, hospitalizations) associated with short-term variations in pollution levels, and chronic exposure studies which investigate the effects of low or moderate exposures that persist for long periods as well as the cumulative effects of repeated exposure to substantially elevated pollution levels (Pope III, 2000).

Epidemiological studies on the effects of particulate air pollution on health found that short-term exposure to high particle concentrations can be linked to increased mortality, increased hospital admissions and reduced lung function as well as aggravated pre-existing conditions such as asthma and cardiac disorders. Long-term exposure can lead to increased mortality and respiratory diseases but seems to be more closely linked to cardiovascular diseases. Particularly vulnerable groups of the population, such as the elderly or chronic sick, are more at risk. It is generally understood that there is no lower threshold for particulate pollution health effects, i.e. even at low particle concentrations, health effects still exist (e.g. COMEAP, 1998; Pope III, 2000; Weijers et al., 2001).

The Committee on the Medical Aspects of Air Pollutants (COMEAP) has evaluated research from the US, and more recently Europe, and found that, although the results can only be transferred to the UK in a qualitative sense, it would be prudent to consider the associations between exposure to particulate air pollution and health as causal (COMEAP, 1995, 2001). They also conclude that, since there is no strong evidence on the relative effects of different particles in the respirable range, the measurement of $PM_{10}$ for policy purposes is reasonable. Other studies, however, come to the conclusion that $PM_{10}$ and even $PM_{2.5}$ do not represent the particles having the most detrimental effect on health, and recommend to focus on smaller size fractions and number concentration measurements (i.e. ultrafines) (Weijers et al.,

2001; Pope III, 2000; Wilson et al., 1997, Keywood et al., 1999). In fact, Wilson et al. (1997) suggest that coarse and fine particle fractions should be investigated separately due to the large differences in their origin, composition and properties. As discussed above, this is supported by many studies (e.g. Molnár et al., 2002; Harrison et al., 1999). The agreement of health studies and empirical air quality research on this issue emphasises its importance.

The findings that ultrafine particle concentrations are especially high next to the road (e.g. Hitchins et al., 2000; Wehner et al., 2002) suggest that people who spend a significant amount of time on or near roads (e.g. car commuters) are exposed to particularly high concentrations of small particles. A number of studies have confirmed this (e.g. Harrison et al., 1999; Jansen et al., 2001; Weijers et al., 2001), and commuter studies have shown that particle concentrations can reach even higher values inside vehicles (Morton, 2002; Gee et al., 1999; Adams et al., 2001a; Kingham et al., 1998).

It is generally agreed that knowledge about particulate matter air pollution is currently limited and more research is needed in order to identify the underlying biological mechanisms and the relation between exposure and health effects.

## 2.3 Personal exposure

As noted above, the mechanisms of air pollutant dispersion in urban areas are rather complex, which makes it problematic to use modelling approaches to estimate concentrations of those pollutants and to assess how they affect individuals or population groups. Exposure research therefore tends to focus on pollutant measurements in conjunction with information regarding the location and duration of people's activities, so called time activity data. This approach allows an insight into the spatial and temporal variability of concentration levels that individuals are exposed to.

This section gives a brief description of the theoretical background of personal exposure research, discusses some issues regarding the inference of personal and population exposure from concentration measurements and provides an overview of personal exposure studies carried out in transport microenvironments.

### 2.3.1 Definition

Although monitoring concentrations of air pollutants provides valuable data about the air quality in an area, it does not, on its own, enable an assessment of to what extent humans passing through that area are exposed to pollutants. In order to evaluate the levels of air pollution a person experiences, account has also to be taken of how long they are in contact with the pollutant(s). It is important to make a distinction between the terms concentration, exposure and dose. Concentration is defined as a physical characteristic of the environment at a certain place and time, whereas exposure refers to any contact between an airborne pollutant and a surface of the human body. Dose is the amount of the pollutant that actually enters the body and reaches the target tissues. The dose of particulate matter therefore varies depending on how fast and deep a person breathes (Hertel et al., 2001; Sexton and Ryan, 1988).

Sexton and Ryan (1988) define personal exposure as an individual's contact with a pollutant concentration which takes place when the pollution concentration at a particular time and place is non zero and the person is present in that same place and at that time. The dimensions of exposure are *mass*time/volume* or *concentration * time*, e.g. $\mu g\ m^{-3}\ h$.

Personal exposure can be determined by direct or indirect methods. The direct method relies on portable personal exposure monitors or measurement of biological markers. Indirect methods use measured concentration data and information about time spent in specific microenvironments in order to determine exposure (Hertel et al., 2001, NRC, 1991a, 1991b). In Sexton and Ryan (1988) a microenvironment is defined as a three-dimensional space where the pollutant level at some specific time is uniform or has constant statistical properties. Using the indirect approach, a general equation (Eq 2-1) for the calculation of integrated exposure of various people to multiple contaminants in diverse microenvironments is given by NRC (1991a) as

$$E_{i,j,k} = \sum_{i=1}^{I} C_{i,j} \cdot \Delta t_{j,k} \qquad \textbf{(2-1)}$$

where $E_{i,j,k}$ (e.g. $\mu g\ m^{-3}\ h$) is the exposure of the $k^{th}$ person to the concentration $C_{ij}$ of the $i^{th}$ pollutant in the $j^{th}$ microenvironment for the time $\Delta t$. A constraint of the indirect approach is that so called time-activity pattern information is required, which

indicates how much time people spend in different microenvironments. However, if only the exposure of one person to a certain pollutant in a particular microenvironment is investigated, such information is not necessary and Eq 2-1 can be simplified. Taking into account that the concentration varies over time, the exposure $E$ can be expressed as follows

$$E = \int_{t_1}^{t_2} C(t)dt \qquad \textbf{(2-2)}$$

where $C(t)$ represents the functional relationship of concentration with time for an interval $t_1$ to $t_2$ (NRC, 1991a). An operational form of the above equation, which allows the calculation of exposure from measured concentration data, is given by

$$E = \sum_{i=1}^{I} C_i(\Delta t_i) \cdot \Delta t_i \qquad \textbf{(2-3)}$$

where $C_i(\Delta t_i)$ is the average concentration to which the person is exposed during time interval $\Delta t_i$.

This exposure value $E$ represents an 'integrated exposure'. However, a more commonly used measure is 'average exposure', which is calculated by dividing the integrated exposure by the time during which the integrated exposure was experienced and has the unit *mass/volume*, i.e. the unit of concentration. In air quality research the average exposure is thus often calculated as the average of the measured pollutant concentration and the terms 'average concentration' and 'average exposure' are used interchangeably.

Modern monitoring equipment allows the measurement of particle concentrations for short time intervals, i.e. down to 1 second. This makes it possible to investigate the concentration of pollutants, and hence exposure, at a finer resolution so that peaks can be observed and possible sources identified more easily.

As pointed out by Oglesby et al. (2000), personal exposure studies are demanding for participants who often have to carry personal monitoring devices and record time activity data. Participants are aware of being monitored and may thus modify their time-activity patterns during measurements ("Hawthorne Effect"). By investigating personal exposure in one particular microenvironment, i.e. evading the use of time-

activity data, these effects do not have to be taken into account, which somewhat simplifies the study set-up.

### 2.3.2 Personal exposure vs. ambient, indoor and outdoor concentrations

In order to estimate group or population exposure to air pollutants, it would be ideal to have personal monitoring data available. However, large scale personal exposure monitoring is usually infeasible due to the large number of monitors and volunteers required. Also, problems may arise when generalizing results from individuals to the population of interest, since time-activity information is not usually available for large population groups. Therefore, a number of studies have investigated the possibility of inferring population exposure using data from fixed site monitoring stations. However, the relationship between ambient concentrations of particulate matter (PM) and personal exposure to PM appears to be more complex than initially expected. For instance, Oglesby et al. (2000) report that they found strong correlations between home outdoor $PM_{2.5}$ concentrations (i.e. ambient concentrations near the home) and fixed-site $PM_{4.0}$ concentrations but no correlation between personal exposure and either of those concentrations. However, epidemiological studies have shown that statistical associations do exist between levels of particle concentrations and health endpoints (e.g. hospital admission, mortality etc.). In fact, these statistical relationships form the basis for current air quality standards for particulate matter.

Morawska et al. (1999) suggest that the problem lies in the large spatial variation of concentrations of $PM_{10}$ and $PM_{2.5}$, which tend to be higher near roads. Values from fixed site monitors can therefore not be considered to represent evenly distributed concentrations in a large target area. They recommend that, since the identified correlations were weak but significant, the use of ambient PM concentration values may have merit if relating the studied population to the area in which they live, i.e. use monitoring data from stations as close to homes as possible. Keywood et al. (1999), who investigated relationships between size segregated mass concentration data, came to the conclusion that, since variation in $PM_{10}$ is dominated by variance in the more dangerous $PM_{2.5}$ fraction, an increase in $PM_{2.5}$ monitoring may aid the detection of a relationship between health effects and ambient particle concentrations.

Wilson et al. (2000) refer to the conflict of reported health effects despite the absence of any link between ambient concentrations and personal exposure as the 'exposure paradox' and explain that, in addition to outdoor ambient sources, the effect of indoor sources needs to be considered when investigating personal exposure to particles. They differentiate between ambient-generated PM (ambient outdoor PM plus ambient PM that has infiltrated indoors) and indoor PM (generated entirely from indoor sources). They also use personal activity PM as an additional category to define microscale particle-generating activities that primarily affect the person performing the activity. This is a wider definition than the 'personal cloud' one which is defined by Wilson et al. (2000) as the difference between a personal monitor measurement and an area-representative measurement a few metres away. The approach used by Wilson et al. (2000) allows a separation of indoor and outdoor PM depending on their sources rather than where they were measured. When applied to the microenvironment 'in-vehicle' this means that, due to the lack of indoor sources, measured particle concentrations must originate from either outdoor or personal activity sources. Considering that dust particles are present within the car, which may get re-suspended by movements of the driver, tend to belong to the larger size fractions ($PM_{10}$ or greater), most of the measured $PM_{2.5}$ and all of the ultrafine particles can be assumed to have originated outside the vehicle. They will therefore depend on outside concentrations. With the monitors placed in the confined drivers cabin, personal exposure measurements will include the effect of the personal cloud.

A number of studies comparing outdoor and in-vehicle concentrations have shown that particle concentrations inside a vehicle often exceed outside concentrations (Morton, 2002; Gee et al., 1999; Adams et al., 2001a; Kingham et al., 1998). However, the exact relationship between outside and inside values has not been identified. Clarke (1998), who investigated driver exposure to CO, likened the effect of the ventilation system of a vehicle to the dynamic response of a low pass filter, removing pollution peaks of short duration. Whether this applies to particle concentrations is not clear. Clarke (1998) also cites further research that found that the air exchange rate of stationary vehicles varies from 1 to 3.7 $h^{-1}$ depending on window and ventilation settings. This indicates that there will not be an instantaneous rise in inside concentrations when concentrations increase outside the vehicle, but rather a delayed response is to be expected.

Dickens (2000) states that in-vehicle PM concentrations are lowest with the ventilation switched off, whereas Rhodes et al. (1999) come to the conclusion that vehicle type and ventilation settings have little effect on the in-vehicle concentration of pollutants. A recent study by Chan and Chung (2003) investigated the indoor-outdoor (IO) ratio for pollutants in cars for different ventilation settings, road types and driving conditions. They found that the IO ratio can vary significantly for one ventilation mode (e.g. air-conditioned, windows open) when travelling in different driving environments. Chan and Chung (2003) conclude that no consistent pattern of IO characteristics could be observed due to the many other interacting factors involved. They also report that the correlation of indoor to outdoor values was small for NO and $NO_x$ but exceptionally high for CO. They suggest that this is due to the different pollutants showing different penetration patterns. A study by Chang et al. (2000), who investigated correlations between personal and ambient concentrations of $PM_{2.5}$, found that the strongest correlations were found for microenvironments with high air exchange rates, i.e. in vehicles. They report that stronger correlations were observed for winter data than for summer data ($r^2$ = 0.90 and 0.69, respectively). Since these results were based on hourly averages, it is therefore unclear to what extent they apply to the infiltration of particulate matter into vehicles on shorter time-scales.

However, it is very likely that the 'response' of in-vehicle particle concentrations does depend on speed, the type of vehicle used (i.e. its air tightness) and the ventilation settings. It is therefore important to make vehicle specific measurements when investigating the indoor-outdoor relationship of pollutants in driver exposure studies.

### 2.3.3 Transport exposure studies

Since people generally do not live near their work place but tend to commute, they spend an a significant amount of time in traffic when travelling to work or shopping, i.e. next to one of the main sources of pollutants. A number of studies have been carried out which investigated the exposure of commuters to certain air pollutants in order to quantify exposure levels and to identify differences in travel mode (e.g. bus, bicycle, car). Table 2-2 gives an overview of personal exposure studies which included in-vehicle measurements of particle concentrations.

| Type of study | Location | Transport modes | Pollutants measured | Data resolution/ measuring interval | Reference |
|---|---|---|---|---|---|
| Fixed route commuter study | London, UK | Bicycle, bus, car, under-ground | $PM_{2.5}$ | Average journey concentrations | Adams et al., 2001a and 2001b |
| Fixed route commuter study | London, UK | Bicycle, bus, car, under-ground | Elemental carbon (as an indicator of diesel exhaust) | Average journey concentrations | Adams et al., 2002 |
| Fixed route study | Kuopio, Finland | Car | Particle counts (0.3 - 25 μm) and size distribution<br>CO | 1 min | Alm et al., 1999 |
| Multiple micro-environment study | Baltimore, MD, USA | Mini-van, bus | $PM_{2.5}$<br>CO, $O_3$, $NO_2$, $SO_2$, VOC | Hourly averages | Chang et al., 2000 |
| Rural/urban route comparison | Calham, London, UK | Two different cars | Ultrafines (?)<br>Particle mass<br>$NO_2$, CO<br>Particle size | 1 sec<br>5 sec<br>5 sec<br>180 sec | Dickens, 2000 |
| All day personal monitoring and fixed route study | Connecticut, USA | Bus | $PM_{10}$, $PM_{2.5}$ | 10 sec | EHHI, 2002 |
| Commuter study (various routes) | Manchester, UK | Car, bus | Respirable (~$PM_4$) | Weekly average | Gee et al., 1999 |
| Fixed route study | Mexico City | Bus, minibus, Metro | $PM_{2.5}$, total carbon (TC)<br>CO | Average journey concentrations<br>1 min | Gómez-Perales et al., 2004 |
| Fixed route study | Northampton, UK | Car, pedestrian | $PM_{10}$, $PM_{2.5}$, $PM_1$ | Average journey concentrations | Gulliver and Briggs, 2004 |
| Commuter study (comparison of modes, one route) | Marsden, Hudders-field, UK | Train, bus, car, bicycle | Benzene<br>Particles | ?<br>Absorbance | Kingham et al., 1998 |
| Driver exposure | Bradford, UK | Car (stationary) | Particle counts (1 - 15 μm) | 1 min | Powell, 2000 |
| Variety of scenarios (total of 27 runs) | California, US | Car | $PM_{2.5}$, $PM_{10}$<br>Particle counts (various size ranges), CO | Integrated mass conc.<br>1 min and 120 min | Rhodes et al., 1998 |

**Table 2-2 Personal exposure studies which investigated in-vehicle exposure to particulate matter**

Such travel mode studies address an important exposure issue. As explained in Colvile et al. (2001), time averaged concentrations can vary significantly within a few metres of or immediately next to a road, which means that road users may be affected by rather different levels of pollution depending on their exact location. As noted above, particle concentrations on roads are influenced by localised variations in atmospheric conditions in conjunction with emissions from individual vehicles as well as traffic flow conditions and road side buildings, and are therefore very difficult to model. Thus, the empirical approach of personal exposure measurement with portable monitors is often used in order to derive a general idea of concentration levels and variability of personal exposure to particles for different travel modes.

Most of these studies used integrated pollution measurements, i.e. determined hourly averages from total concentration or exposure values measured over the whole of a journey. However, it has been shown by Dickens (2000) that the exposure to pollutants does not only depend on travel mode but also varies significantly for different types of road layout and location (i.e. motorway, small urban road etc.) and particularly vehicle speed. The difficulty in determining commuter or driver exposure is that it does not solely depend on pollutant concentration inside the vehicle, but is also a function of journey time and speed. On the one hand, vehicles emit more particles when travelling at higher speeds, which causes an elevation of ambient levels. On the other hand, the journey time and hence potential exposure time is reduced due to higher speeds. This means that, for instance, the actual measured concentrations inside the car on motorways may be relatively high compared to slow moving traffic in the city centre, but because the time spent on motorways may be shorter due to higher speeds, the exposure to pollutants on the motorway could be lower (Dickens, 2000).

Personal exposure studies, such as the ones mentioned above, give an indication of the exposure of individuals on particular journeys but, apart from Dickens (2000), cannot account for exposure variation on different parts of the journey. Findings from these studies can therefore find only limited use in estimating exposure for other journeys. In order to provide such estimates, something like an 'exposure factor' would be needed which, depending on certain parameters such as road layout and typical speed data, would give an indication of the exposure to pollutants that an individual would experience when travelling along a particular type of road. If such

'exposure factors' existed for each type of road (i.e. dual carriage way, residential road etc.) any journey could be split up into a number of 'links' which, according to their characteristics, would be attributed an appropriate 'exposure factor'. Using journey time information and vehicle flow data, exposure values could be calculated for each link. Results for the road links could then be summed to provide an overall exposure estimate for the journey.

One of the objectives of the research presented in this thesis was to investigate whether this hypothesis is valid, i.e. whether road layout and dynamic parameters can be shown to affect driver exposure levels (Chapters 4 and 5). The potential of the findings for exposure modelling is also explored (Chapter 6).

Chapter

# 3 Field Campaign: Data Collection and Processing

The objective of the work presented in this thesis was to investigate personal exposure in a transport microenvironment. The study required in-vehicle particle concentration data for ultrafine particles collected on different types of roads. As noted in Section 2.3.3, data from other exposure studies was not suitable. It was thus necessary to carry out a specially designed data collection exercise in order to acquire a reliable data set as a basis for a comprehensive analysis.

This chapter describes the data collection methodology which was developed for the this study and applied during the field campaign, including equipment details and route choice. The chapter also outlines the procedures applied to prepare the data for the analysis, such as data screening and pre-processing, and describes the final field study database. Due to the large size of the obtained field data set, the various procedures required a significant amount of automation which was achieved by using custom written programming routines, the main aspects of which are the focus of the final section of this chapter.

## 3.1 Field study

The field study described here formed the basis for two PhD projects; one investigating driver exposure (this thesis) and the other regarding drive cycle research

(Turpin, 2004). Although these were two distinct studies, they were able to share a data collection platform. The two authors were jointly responsible for the development of the common parts of the field study methodology (i.e. division of routes into links, set-up and calibration of the CO and speed sensors) and the execution of the data collection campaign (i.e. driving the data collection vehicle). The choice of routes, the data collection schedules and the particle monitoring equipment, however, was the sole responsibility of the author of this thesis, as was the development and implementation of the data processing methodology.

Prior to the data collection phase, requirements were defined, equipment options were investigated and a suitable vehicle was configured as a mobile monitoring unit. Based on information from the literature review and a number of trial runs, a methodology for the data collection process was developed and refined.

### 3.1.1 Requirements

As the field study described here formed the basis for two projects, it needed to meet a wide range of requirements. The data collection methodology was developed based on the following considerations:

Which parameters?

As discussed in Section 2.2.3, ultrafine particles ($< 0.1\ \mu m$) are suspected to more strongly affect human health than the more commonly measured larger particles (i.e. $PM_{10}$ and $PM_{2.5}$). Study results also indicate that particles of the different size ranges show dissimilar spatial and temporal behaviour, which suggests that mass concentration measurements cannot be used as surrogates for the smaller particle size fractions (discussed in Section 2.2.1). It was therefore considered important that mass and number concentrations were measured simultaneously during the field campaign in order to gain an understanding of the nature of particle exposure in vehicles in general and to investigate the differences and similarities in the behaviour of the different particle size fractions.

Instantaneous speed of the vehicle was to be collected for the drive cycle study (Turpin, 2004). It was envisaged that these data would also allow an investigation of the relationship between vehicle speed and in-vehicle particle concentrations.

Carbon monoxide (CO) concentrations, which can be used as a marker for traffic emissions, were to be logged as an inexpensive way of acquiring information on an additional pollutant, which may be useful in the analysis phases of both studies or for a separate subsequent investigations.

In order to be able to use the data to investigate the temporal dynamics of driver exposure, it was considered important that the monitoring equipment facilitated high frequency measurements, i.e. logging data at short time intervals of a few seconds.

Where?

Since in-vehicle particle exposure represents a problem mainly in urban areas where high traffic density is observed, the data collection was to focus on locations in an urban centre (i.e. the City of Leicester). The selected areas needed to be easily accessible from the location of the Institute and include various typical urban types of road, from dual carriage ways to residential roads, to allow the investigation of driver exposure depending on road categories.

Different traffic flow patterns are observed during morning and evening rush hour, with the majority of vehicles going towards the city centre in the morning and away in the evening. In order to cover all situations, data was to be collected on all selected roads for both directions (where possible).

Other practical considerations included choosing the field study location(s) so that they could be reached easily (i.e. near to where the Institute is based) and data collection could be completed within the peak hour.

When?

Regular vehicle commuters are responsible for the majority of high-density traffic, which generally has a morning and early-evening peak, i.e. rush hour (Harrison et al.,1999a). These commuters spend the longest time in the vehicle microenvironment when the pollutant levels are likely to be the greatest because of the high traffic densities. It was therefore considered appropriate to schedule data collection runs to occur during the peak traffic periods of 8 - 9 am and 5 - 6 pm.

It is understood that meteorological conditions have a significant influence on particle concentrations (see Section 2.2.1). In order to account for this seasonal variability, the data collection runs were scheduled to cover a whole year.

How?

The data collection required a mobile monitoring unit which could be equipped with all necessary monitoring devices. At the same time it needed to be a 'typical' vehicle in order for the results to be as representative as possible for normal driver exposure conditions. It was considered desirable to use an electric vehicle which would not contribute to traffic emissions and thus support the institute's endorsement of sustainable development.

To be able to relate in-vehicle particle concentrations to traffic flow patterns and events (e.g. queuing), it was necessary to record instantaneous speed of the vehicle as well as general information about traffic events (e.g. location on route, which car is being followed).

The field campaign was to be carried out using two drivers (i.e. the author and the drive cycle researcher), who would alternate on a daily basis.

In order to devise a field study which would take account of all these considerations, various equipment, route and methodology options were investigated. A number of trial runs were carried out to test and refine the initially proposed methodology and to familiarize the drivers with the routes, the equipment and the recording procedures. The following sections describe the main aspects of the final study setup and methodology which was used throughout the field campaign.

### 3.1.2 Instrumentation

An electric van (Peugeot Partner) was used as a mobile monitoring unit. The vehicle is based on a typical hatchback chassis but powered by an electric motor. The van has not been modified in any other way, and with regard to ventilation and other parameters that may affect pollutant intake, it is considered equivalent to a typical domestic hatchback[1]. It is acknowledged that the use of an electric vehicle may result in an underestimation of actual driver exposure values for vehicles with combustion engines, which are known to be partly influenced by the vehicle's own exhaust (e.g. Behrentz et al., 2004). However, the contribution of self-pollution is likely to vary for different vehicles and driving conditions. The use of an electric vehicle, which does not contribute to local ultrafine particle emissions, is therefore considered

1. Peugeot - Private communication.

advantageous since it allows the investigation of driver exposure to emissions from the surrounding traffic only, i.e. by removing the extra variability in particle concentrations resulting from self-pollution.

The vehicle was equipped with extra batteries to power the on-board equipment, which consisted of two particle monitors for different size ranges, a CO sensor and a speed sensor.

OSIRIS environmental dust monitor

The OSIRIS (Turnkey Instruments Ltd.) is an optical device which can measure four size ranges of fine and coarse particles simultaneously (i.e. total particle mass, $PM_{10}$, $PM_{2.5}$ and $PM_1$ equivalent). The particles are detected by light scattering, from which the size of each particle is inferred. Using the size information, the measured number concentrations are converted into mass concentration values by using a pre-defined mass calibration factor. The monitor has a flowrate of 600 $cm^3$ $min^{-1}$ and can detect particles in the size range from 0.4 to 20 μm (Turnkey Instruments, 2001).

Due to the optical detection method used, the particle concentration values measured with the OSIRIS can only be considered estimates of $PM_{10}$ and $PM_{2.5}$ mass concentrations. However, as noted in Section 2.2.2, Gulliver and Briggs (2004) reported that OSIRIS measurements were well correlated with TEOM data and generally only slightly exceeded them. Based on these observations and the fact that the monitor has been calibrated by the manufacturer against a $PM_{10}$ reference concentration, the measured values are considered reasonably accurate estimates of the standard mass concentration metrics and are thus referred to as $PM_{10}$ and $PM_{2.5}$ concentration values throughout this thesis.

P-Trak ultrafine particle counter

The P-Trak (model 8525, TSI Inc.) is an ultrafine condensation particle counter which can detect particles with sizes ranging from 0.02 μm to greater than 1 μm and has a flowrate of 100 $cm^3$ $min^{-1}$ (TSI, 2003).

The number concentration values measured with the P-Trak thus include particles outside the ultrafine range (i.e. > 0.1 μm). However, as discussed in Section 2.2.1, particles above 0.1 μm do not significantly contribute to number concentrations and, for example, findings by Junker et al. (2000) have shown that, in rush hour traffic, ultrafine particles contribute up to 97 % to the total particle number (< 0.421 μm).

The data measured with the P-Trak are thus referred to as 'ultrafine particles' or 'UF' throughout this thesis. However, it is important to note that the P-Trak may erroneously record changes in particle concentration when the particle size distribution changes around the detection limit of 0.02 μm , e.g. caused by variations in aerosol composition and/or meteorological conditions.

Carbon monoxide sensor

Carbon monoxide was measured using an electro-chemically based sensor (CiTicel 3E/F, CityTech Ltd.) whose output was logged with a TinyTag*Plus* data logger (TGPR-0804 Gemini Data Loggers (UK) Ltd.). The sensor arrangement was installed in a tool box so that it could be used as a portable monitor. Note, the CO data acquired during the field study was not included in the subsequent analysis of the field data, as presented here, but may be utilized for further separate data analysis.

Speed sensor

Another TinyTag*Plus* data logger (same type as above) was used to log data from the frequency-analogue converter which was connected to the on-board speedometer. This current data (mA), representing instantaneous speed, was converted to miles/hour (mph) during data processing. The conversion function was calibrated based on calibration data obtained during a dynamometer test at the MIRA test facility, as described in Appendix A.1. Note, the installation and calibration of the speed sensor was carried out in close collaboration with the driver cycle researcher.

Installation and logging times

The OSIRIS monitor was installed at head height on the bulkhead behind the passenger seat. Both, the P-Trak monitor and the CO sensor box where placed on the passenger seat and secured using the seat belt. The data logger for the speed sensor was installed semi-permanently on the bulkhead to facilitate data download.

The equipment was powered by two batteries with inverter (located in the back of the vehicle), which provided mains power, while the CO sensor arrangement contained its own 12 V battery.

The monitoring equipment allows logging of data at time intervals down to 1 second (1 Hz). Initial test runs revealed that the number concentration of ultrafine particles can vary significantly over periods of just a few seconds, and that this can be related to specific events such as a transient close approach to a car directly in front. It was

decided therefore to record all parameters at 1 Hz and maintain them in the database at this frequency. All subsequent analysis would be based on the 1 second, or averaged over longer periods if found necessary.

However, it is important to note that both the air change rate within the instruments and their electronic response time could affect the interpretation of short-term variations in pollutant concentrations.

Quality assurance

The OSIRIS particle monitor was calibrated against a $PM_{10}$ reference instrument[2] by the manufacturer prior to the field campaign and subsequently at annual servicing intervals specified by the standard maintenance schedule. Routine flowrate checks were carried out regularly as recommended by the manufacturer (Turnkey Instruments, 2001).

The P-Trak monitor was calibrated against a CPC type 3010 (TSI)[3] by the supplier prior to the field study campaign and subsequently at annual servicing intervals specified by the standard maintenance schedule. A daily 'zero check' was carried out, using the zero filter supplied with the instrument, in order to verify that the instrument and alcohol cartridge were assembled properly and free from leaks (TSI, 2003).

The CO sensor arrangement was calibrated against a reference gas twice during the field campaign and the speed sensor assembly was calibrated prior to the data collection with data from a dynamometer test as described above. Although, as noted, the CO data are not used in this analysis, they were logged together with the other data channels throughout the field campaign.

### 3.1.3 Location

The field study was carried out in Leicester, which is the largest city in the East Midlands with a population of 280,000. The main industries are manufacturing, retailing and services (Census, 2001).

2. Accuracy within ±10% of reference instrument, which is itself calibrated annually against reference values from gravimetric measurements (Turnkey Instruments, 2001).
3. Details included on the calibration certificate provided by the supplier.

Three field study routes were devised which are located within 2.5 km of the city centre (see Figure 3-1). The routes were devised so that they included different road types, i.e. dual carriage way, residential streets, one way system with bus lane.

A road link in this context is section of road with constant physical and dynamic parameters. Based on this definition, a new link starts where the traffic flow (i.e. number of vehicles) changes, i.e. due to cars turning off or joining, or where the road layout changes. Both instances typically occur at crossroads and junctions, which results in road links being often terminated by traffic lights or crossroads.

All three routes were circuits (i.e. loops) so as to minimise time loss when subsequent collection runs are carried out on one route. Each route comprised of 11 to 14 road links. Generally, the links are based on the road sections of the traffic flow model TRIPS (Citilabs, 2002) used in Leicester. This allowed traffic flow information from the model, provided by Leicester City Council[4], to be used in the data analysis phase. However, since the model network does not cover the whole of Leicester, some links were added to complete the circuits.

Figure 3-1 provides an overview of the field study routes, showing the locations of the routes and the road links. TRIPS links are marked by green lines and non-TRIPS links in blue. Additional maps showing traffic flow direction and traffic signals on the individual road links are included in Appendix A.2.

Route 1 - Uppingham Road

The beginning of this route (links 1 to 4) forms part of the main access route into the centre of Leicester from the west. These links are mainly dual carriage ways which are strongly affected by rush hour traffic. The other seven links (5 to 11) run through a residential area characterised by single lane streets, bus routes and on-street parking. A primary school, local shops and many small and medium sized businesses are located alongside these links, which are often congested during peak times.

Route 2 - Abbey Lane

This route mainly consists of dual carriage way access routes leading into the centre of Leicester (links 21, 22, 25, 26 and 29 -31) as well as part of the inner ring road (links 26 and 27). The route encompasses the recreation area of Abbey Park and a

4. Modelled traffic flow data was provided by the Pollution Control Group of Leicester City Council.

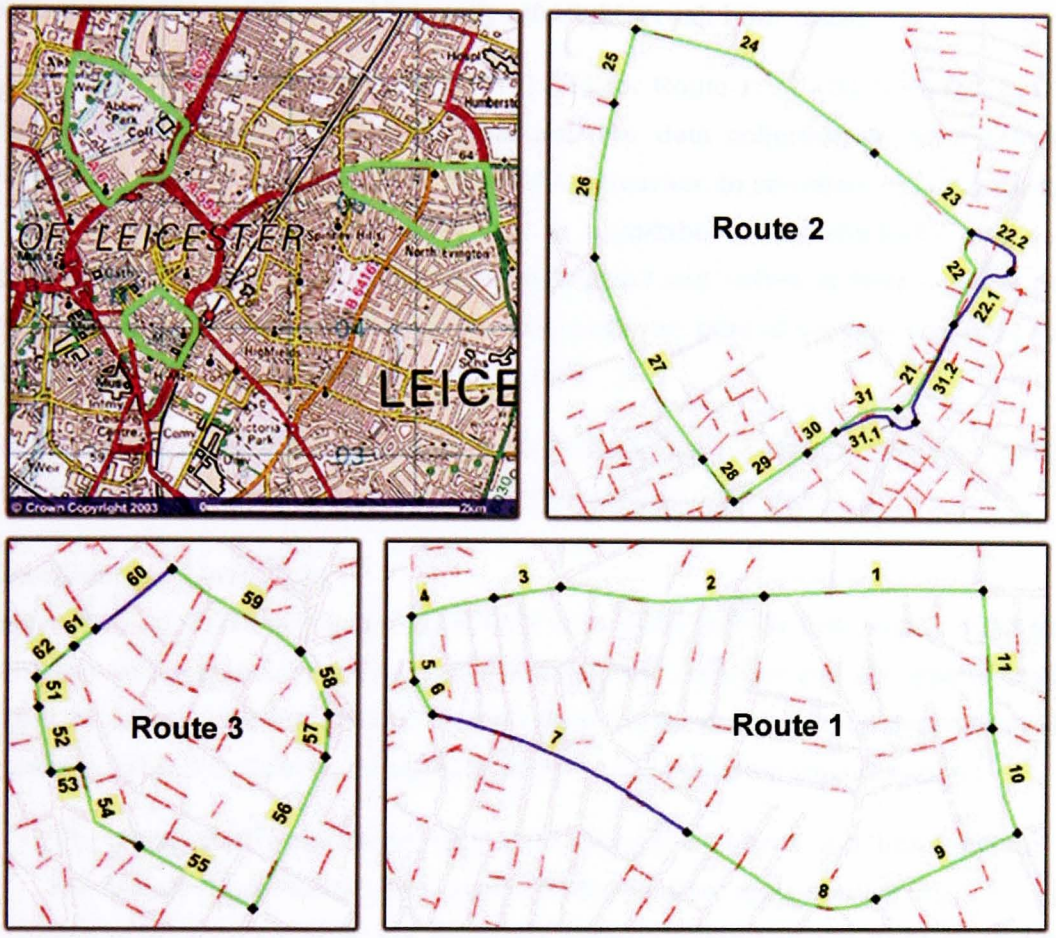


**Figure 3-1 Field study routes**

*clockwise from top left: location of routes in Leicester[a], Route 2 (Abbey Lane), Route 1 (Uppingham Road) and Route 3 (City Centre)[b]*

a. Reproduced from Ordnance Survey map data by permission of Ordnance Survey, © Crown copyright. (green outlines added by author).

b. GIS maps: © Crown copyright. All rights reserved. Leicester City Council. Licence number 100019264 (2004).

small business area which borders onto the ring road. The only residential dwellings on this route are alongside link 24. This route is part of the Leicester bus network and is greatly affected by traffic congestion during morning and evening rush hours.

Route 3 - City Centre

This route follows the one-way system through the very centre of Leicester which has bus traffic and mainly canyon type streets, i.e. narrow streets with high buildings on either side which may limit the dispersion of pollutants. Links 51 and 52 are part of the inner ring road and have a four-lane carriage way layout while links 54 to 57 are

access routes to the centre. The route is busy throughout business hours and the least central links (51 to 57) are additionally affected by rush hour traffic.

Link numbers were allocated as follows: 1 - 11 for Route 1, 21 - 31 for Route 2, 51 - 62 for Route 3 in anti-clockwise direction. For data collection in the clockwise direction, '.1' was added to the link number as a marker. In two cases, this numbering approach was not suitable because, due to roundabouts, the clockwise and anti-clockwise sections of the link had different lengths and varied in their physical and dynamic characteristics. In these cases the clockwise part of the link was split into two parts and the additional link denoted with '.2'.

An overview of all road links included in the field study is provided in Table 3-1, which also shows the physical and dynamical parameters for each link.

Classification of road links

Parameters used to classify the field study road links include link length, road side environment, presence of pedestrian crossings, link function and the type of traffic signal terminating the link (link end). The categories for the individual parameters are shown in Table 3-2. They were defined based on the following considerations.

- The categories for the parameters 'road side environment' and 'link function' are loosely based on the ones typically used in drive cycle studies (e.g. Ericsson, 2001). The former facilitates differentiation between links located in residential and non-residential areas while the latter gives an indication of the main purpose of the road. It indirectly incorporates road size, i.e. number of lanes and capacity, with 'local' and 'high street' being typically smaller than the arterial roads, which are often characterised by a multi-lane layout.
- Links were also classified according to the type of junction layout terminating the link. The categories used for the parameter 'link end description' were chosen in order to best represent the suspected effect of junction layout on waiting times, i.e. longer waiting times were expected on un-signalled right turns while 'open' links should be least affected.
- The parameter 'crossing' was used to classify links in terms of pedestrian crossings, which are likely to have an effect on the traffic flow characteristics.

| | general parameters | | | | anti-clockwise | | | | | clockwise | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | link no. | average length | road side env.[a] | crossing | function[b] | link end[c] | speed limit | trips flow am | trips flow pm | function[b] | link end[c] | speed limit | trips flow am | trips flow pm |
| | | [km] | | | | | [mph] | [# $h^{-1}$] | | | | [mph] | [# $h^{-1}$] | |
| Route 1 | 1 | 0.434 | res. | yes | art. inb. | signal | 30 | 1025 | 838 | art. outb. | r. signal | 30 | 839 | 1168 |
| | 2 | 0.404 | res. | yes | art. inb. | signal | 30 | 1022 | 718 | art. outb. | signal | 30 | 626 | 1108 |
| | 3 | 0.133 | res. | yes | art. inb. | open | 30 | 811 | 822 | art. outb. | signal | 30 | 702 | 1115 |
| | 4 | 0.166 | res. | no | art. inb. | signal | 30 | 968 | 906 | art. outb. | open | 30 | 765 | 1243 |
| | 5 | 0.127 | res. | no | local | open | 30 | 198 | 473 | local | r. turn | 30 | 454 | 206 |
| | 6 | 0.077 | res. | yes | local | open | 30 | 158 | 454 | local | open | 30 | 399 | 184 |
| | 7 | 0.560 | res. | no | high st. | signal | 30 | 40 | 210 | high st. | open | 30 | 163 | 56 |
| | 8 | 0.415 | res. | no | local | open | 30 | 123 | 343 | local | signal | 30 | 562 | 149 |
| | 9 | 0.311 | res. | no | local | signal | 30 | 222 | 550 | local | open | 30 | 769 | 276 |
| | 10 | 0.215 | res. | no | int.-urb. | open | 30 | 643 | 640 | int.-urb. | r. signal | 30 | 602 | 889 |
| | 11 | 0.275 | res. | no | int.-urb. | signal | 30 | 643 | 640 | int.-urb. | open | 30 | 602 | 889 |
| Route 2 | 21* | 0.239 | non-res. | yes | art. outb. | open | 30 | 1282 | 2155 | int.-urb. | signal | 30 | 780 | 660 |
| | 22* | 0.211 | non-res. | yes | int.-urb. | open | 30 | 481 | 1163 | int.-urb. | open | 30 | 992 | 1189 |
| | 23 | 0.319 | non-res. | no | int.-urb. | open | 30 | 1063 | 1121 | int.-urb. | signal | 30 | 1049 | 975 |
| | 24 | 0.690 | res. | yes | int.-urb. | signal | 30 | 917 | 1119 | int.-urb. | open | 30 | 1044 | 751 |
| | 25 | 0.199 | res. | yes | art. inb. | signal | 40 | 1662 | 621 | art. outb. | signal | 40 | 641 | 2021 |
| | 26 | 0.394 | non-res. | yes | art. inb. | signal | 40 | 1727 | 680 | art. outb. | signal | 40 | 688 | 2047 |
| | 27 | 0.581 | non-res. | no | art. inb. | signal | 40 | 2563 | 1204 | art. outb. | signal | 40 | 1116 | 2796 |
| | 28 | 0.134 | non-res. | no | art. inb. | signal | 30 | 2902 | 2043 | art. outb. | signal | 30 | 1943 | 3090 |
| | 29 | 0.220 | non-res. | no | int.-urb. | signal | 30 | 2897 | 2935 | int.-urb. | r.signal | 30 | 2597 | 3394 |
| | 30 | 0.089 | non-res. | no | int.-urb. | open | 30 | 2843 | 3025 | int.-urb. | signal | 30 | 2499 | 2778 |
| | 31* | 0.203 | non-res. | no | int.-urb. | signal | 30 | 1508 | 1510 | int.-urb. | open | 30 | 1070 | 1327 |
| | 32 | 0.273 | - | - | - | - | - | - | - | art. outb. | signal | 30 | 2086 | 1626 |
| Route 3 | 51 | 0.054 | non-res. | no | int.-urb. | signal | 30 | 2530 | 3000 | - | - | - | - | - |
| | 52 | 0.123 | non-res. | no | int.-urb. | l. turn | 30 | 2530 | 3300 | - | - | - | - | - |
| | 53 | 0.057 | non-res. | no | local | open | 30 | 805 | 421 | - | - | - | - | - |
| | 54 | 0.198 | res. | no | local | signal | 30 | 951 | 1207 | - | - | - | - | - |
| | 55 | 0.241 | res. | no | local | signal | 30 | 699 | 705 | - | - | - | - | - |
| | 56 | 0.314 | non-res. | no | int.-urb. | open | 30 | 650 | 1072 | - | - | - | - | - |
| | 57 | 0.077 | non-res. | no | int.-urb. | signal | 30 | 517 | 615 | - | - | - | - | - |
| | 58 | 0.128 | non-res. | no | int.-urb. | l. turn | 30 | 2323 | 1749 | - | - | - | - | - |
| | 59 | 0.282 | non-res. | no | high st. | signal | 30 | 2530 | 3000 | - | - | - | - | - |
| | 60 | 0.182 | non-res. | no | high st. | signal | 30 | 2530 | 3300 | - | - | - | - | - |
| | 61 | 0.051 | non-res. | no | high st. | signal | 30 | 805 | 421 | - | - | - | - | - |
| | 62 | 0.090 | non-res. | yes | high st. | open | 30 | 951 | 1207 | - | - | - | - | - |

**Table 3-1 Main parameters for the road links of the data collection routes. (For details on the classification categories used refer to Table 3-2.)**

a. categories for road side environment: res. - residential, non-res. - non-residential

b. categories for road function: art. inb. - arterial inbound, art. outb. - arterial outbound, high st. - high street, int.-urb. - inter-urban

c. r. - right, l. - left

* link length is different for clockwise and anti-clockwise directions, average link length listed

The length of the road links was derived from GIS road maps[5] by calculating the length of the line objects, while information on the other parameters was collected as part of a pre-study survey of the routes.

| Parameter | Category | Description |
|---|---|---|
| **road side environment (RF)** | residential | mainly residential buildings |
| | non-residential | mainly non-residential buildings |
| **crossing (C)** | yes | one or more pedestrian crossing points (with or without lights) |
| | no | no pedestrian crossings |
| **link (road) function (RF)** | local | small street, mainly used for local access |
| | high street | small street with shops and small local businesses |
| | inter-urban | primary connecting route which is not arterial inbound or outbound |
| | arterial inbound | primary connecting route, multi-carriage way lanes for traffic towards the urban centre |
| | arterial outbound | primary connecting route, multi-carriage way lanes for traffic going away from the urban centre |
| **link end description (LED)** | open | no traffic signal, next link was entered by going straight ahead |
| | left turn | no traffic signal, next link was entered by turning left at crossroads |
| | signal | traffic signal, basic layout without filter lanes |
| | right signal | traffic signal, separately switched for traffic in filter lane for right turning traffic |
| | right turn | no traffic signal, next link was entered by turning right |

**Table 3-2 Road specific parameters and categories used to classify the field study road links**

Further link parameters listed in Table 3-1 include the statutory speed limit (from pre-study survey) and peak traffic flows for morning and evening rush hour, which were derived from the TRIPS traffic flow model.[6]

### 3.1.4 General methods

Floating car method

An efficient way of collecting traffic data such as speed, speed patterns and variation, with a minimum of equipment is the 'floating car method' where an instrumented vehicle 'floats' in the traffic stream, passing only as many vehicles as pass the test vehicle (DfT, 2002). This ensures that the driver's behaviour is largely determined by prevailing traffic conditions rather than personal driving habits. The recorded drive cycles can therefore be considered to be 'typical' for the traffic conditions on the

5. Provided by the Pollution Control Group, Leicester City Council.
6. Estimated traffic flow values (output from TRIPS model) were provided by the Pollution Control Group, Leicester City Council.

studied road sections. This method is often used as the most effective way of gathering unbiased traffic speed data[7].

Drivers

For practical purposes, the data collection runs were carried out by two drivers. Although, this may introduce uncertainties due to variations in driving styles between individuals, these variations were kept to a minimum by using the floating car method.

Ventilation settings

No attempt was made to alter window and ventilation settings during data collection in order to acquire concentration data for each combination of settings. Rather, the 'personal comfort' approach was applied in which temperature and ventilation were adjusted by the driver to his or her requirements. Since the evidence regarding the effect of ventilation settings on in-vehicle concentrations (discussed in Section 2.3.2) was inconclusive, this method was chosen in an attempt to best represent real world conditions, where car drivers adjust ventilation settings according to the ambient weather conditions. It is acknowledged that the air exchange between the vehicle and its immediate surroundings does affect the change of particle concentrations experienced by the driver. However, since the air exchange rate will vary depending on vehicle type, ventilation settings and vehicle speed, it is difficult to quantitatively account for it.

Manual recording of additional data

As part of the instrumentation setup prior to each data collection run, notes regarding the local weather conditions were recorded on a dictaphone. Ambient temperature was measured with a standard thermometer and atmospheric conditions were assessed in broad pre-defined categories (e.g. no wind/windy/very windy, foggy, wet roads etc.). These 'rule-of-thumb' categories provided basic run specific weather information. However, due to the subjective nature of the assessment, these data were disregarded for analytical purposes in favour of more accurate meteorological data provided by Leicester City Council (Section 5.3.2).

7. Recommended in Technical Guidance LAQM.TG(03) (DEFRA, 2003b) and mentioned in minutes of AQEG meeting (AQEG, 2002).

During the journey from the Institute to the field study routes, note was also made of the current vehicle ventilation and heater settings. If the settings were adjusted again during the journey, this was also recorded on the dictaphone.

Throughout the data collection on the routes, a stopwatch, synchronised with the monitors, and the dictaphone were used to record 'event data', such as entering or leaving a link or joining a traffic jam. For practical purposes, the beginning of a link was defined as the half-way point of the junction crossing. Every time a link was entered, the driver read the time at which that point was crossed from the stopwatch and recorded it on a dictaphone, together with the link number. Additionally, the type of vehicle (e.g. car, van, bus) directly in front of the instrumented vehicle and its age (from registration plate) was recorded. While on a link, a note was made of any change regarding the vehicle in front as well as general traffic conditions, such as queuing periods, traffic congestion in the opposite direction or encountering a particularly polluting vehicle. Each instance was recorded together with the time of occurrence. Monitor malfunction (e.g. 'low alcohol' alert on the P-Trak) were recorded in the same manner.

### 3.1.5 Data collection

Following a number of test runs, which served to familiarize the author (and the other driver) with the vehicle and instrumentation set-up, the collection of particle concentration data started on 2 February 2002, and extended until 9 April 2003. However, due to supplier problems only the OSIRIS monitor and the CO and speed data loggers were initially available. The collection of particle number concentrations started on 24 April 2002. Data collection sessions were carried out on two days a week during morning and evening peak times, i.e. morning rush hour 8 - 9 am and evening rush hour 5 - 6 pm.

Prior to each data collection session, the on-board equipment was set up and checked using the appropriate maintenance procedures, as described in Section 3.1.2. The monitoring devices were then synchronised with each other and a stopwatch, and pre-run notes (date, weather) were recorded. After the setup procedures had been completed, the vehicle was driven to the start of the first scheduled route and data collection commenced, typically at 8 am or 5 pm respectively. Occasionally, traffic conditions would cause a slight delay in starting time. However, this did not represent

a problem since the journey start time was recorded on the dictaphone in any case in order to pinpoint the start of the data collection run within the recorded time series data. Typically, all three routes were visited during the peak hour, in varying order. Routes 1 and 2 were each completed twice during one session, clockwise on one collection day and counter clockwise on the next, while Route 3 is a one-way system which can only be followed in one direction. Occasionally, traffic congestion events resulted in long delays, so that data collection could only be completed for two routes within the peak hour. In order to balance the number of runs completed for each route, the journey schedule was adjusted weekly.

While the instrumented vehicle was driven along the pre-defined routes, in-vehicle concentrations of particulate matter ($PM_{10}$, $PM_{2.5}$, ultrafine), CO and instantaneous speed were recorded. All data were logged at a rate of 1 Hz. Throughout the data collection session, the driver recorded event information on a dictaphone, using a stopwatch to ensure post-run synchronisation with the monitored data, as explained above.

At the end of each session, the data from the four monitors were downloaded to a laptop so that they could be backed up and transferred to the main computer for processing, while event data from the dictaphone was manually entered into a spreadsheet. Data pre-processing and screening were carried out weekly. The procedures used are explained in Sections 3.1.6 and 3.2.

### 3.1.6 Data screening and quality control

After each data collection session, the validity of the four data sets was assessed based on the recorded event data and visual inspection of the time-series data. Events which necessitated the discarding of the associated data were occasionally recorded, such as:

- Initialisation problems that affected data synchronisation (operator error).
- Dictaphone or stopwatch malfunction.
- 'Low alcohol' alert (P-Trak data).
- Exceedance of the concentration range (P-Trak data).
- Traffic accidents which severely disrupted traffic flow.

Visual inspection of time-series data (plots and spreadsheets) were used to detect additional monitor malfunctions. For almost all the sessions no malfunctions were found. Only five instances of missing CO data occurred, which were found to be due to a power failure in the equipment. Other errors observed included sudden changes in CO concentrations which could also be attributed to equipment malfunction.

All data which were suspected to be unreliable, as well as missing data, were replaced in the data file with a marker value (i.e. '-999') so that they could be easily identified during further processing and analysis. In severe cases, where only few valid data remained after screening, the complete run data set was discarded.

It is acknowledged that the field study data is subject to errors and inaccuracies, introduced by the study design and methodology, which are difficult to quantify. The main considerations regarding systematic errors intrinsic to the field study design are:

- Instrument inaccuracies were minimized by regular calibration following the recommended procedures.
- Since only one instrument of each type was used, between-instrument variability could be eliminated as a source of error.
- Differences in driving styles between the two drivers were minimized by using the 'floating-car' method.
- Inaccuracies due to the 'stopwatch-dictaphone' method used to record link start and end points were minimized by using test runs during which the drivers inspected the routes and established a consistent methodology (e.g. half-way points define link start).

A large number of runs were carried out covering a whole year. This ensured that data was collected for a wide range of traffic conditions (i.e. congested, normal, free flow) and meteorological conditions (i.e. dry/wet, warm/cold, windy/calm), so that the influence of extreme events on the analysis results was reduced.

This field study was designed to investigate whether particle concentrations measured inside the vehicle can be related to traffic conditions on the road link they were recorded on. However, it is acknowledged that the air exchange between the outside air and the driver's compartment of the vehicle is not instantaneous, which may result

in particle concentrations measured on a particular link being affected by traffic conditions on the previous link. This issue is especially relevant for the interpretation of data from short links.

### 3.1.7 The field data set

The field study lasted approximately 12 months, during which a total of 133 valid sessions were completed, i.e. journeys for which valid data was available after data screening. During the first half of the campaign, 91 collection sessions were completed. However, after October 2002 only one driver was available. Due to the reduced data collection efficiency caused by this and additional constraints resulting from extensive road works carried out on Route 1 (Uppingham Road) during autumn and winter 2002, slightly more data runs were carried out during the summer months (April - September) than the winter months (October - March). This is illustrated in the bar chart in Figure 3-2. In addition, Table 3-3 provides an overview of the data set, showing the number of data collection runs for which valid data was available, disaggregated for route, season, time of day and direction of travel. A 'run' typically consists of data from two consecutively completed circuits on the same route during the same am or pm session.

| | | Summer | | Winter | |
|---|---|---|---|---|---|
| | | am | pm | am | pm |
| Route 1 | clockwise | 16 | 20 | 8 | 8 |
| | anti-clockwise | 15 | 16 | 8 | 8 |
| Route 2 | clockwise | 15 | 16 | 13 | 13 |
| | anti-clockwise | 16 | 16 | 13 | 12 |
| Route 3 | anti-clockwise | 33 | 32 | 16 | 14 |

**Table 3-3 Data collection runs carried out during the field study, grouped by location, direction of travel, season and time of day**

Approximately twice as many runs were completed during the summer period on Route 3 in anti-clockwise direction than for any of the other locations. This is due all routes being visited an equal number of times during that period and Route 3 being part of a one-way system, which could only be followed in anti-clockwise direction. There is generally less winter data available than summer data, particularly for Route 1, due to the various reasons explained above. A more detailed description of the final field study data set, including descriptive statistics, is given in Section 4.2.

The data collection methodology developed for this driver exposure study was applied successfully during a long field campaign as well as during an additional field study

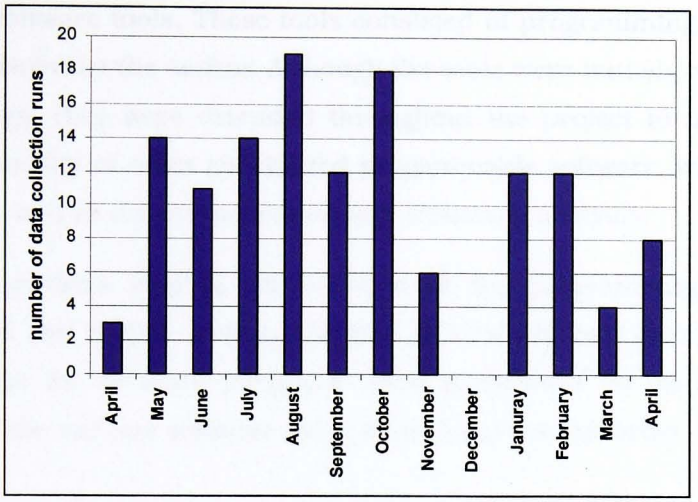


**Figure 3-2 Number of data collection runs carried out per month**

carried out by the Institute of Energy and Sustainable Development for the Leicester City Council as part of their 'Safer Routes to School' project.[8] The methodology aids the collection of in-vehicle particle concentration and instantaneous speed data in combination with event data, which allows external parameters to be linked to vehicle and pollutant dynamics. However, a drawback of the methodology is the manual recording of events, which requires time consuming data input methods and complicates the association of time-series data with event data. An integrated monitoring system incorporating voice recording with automated time-stamp generation would greatly improve the efficiency of the data collection process and should be considered in future studies.

## 3.2 Programming for data processing and analysis

Due to its design and duration, the field campaign generated a large amount of data. An efficient way of managing this data was required in order to aid the post-run processing on a day to day basis and to facilitate data preparation for the analysis. It was immediately apparent at the beginning of the field campaign that this would be a very time consuming and laborious task, which would be prone to operator errors, unless it could be automated by employing programmable software.

8. For this project, an additional route consisting of 13 road links was devised and 19 data collection sessions were carried out by the same two drivers in Spring 2002 (Turpin and Mardaljevic, 2000).

Various software packages were used in combination, which allowed the development of specialized software tools. These tools consisted of programming routines which were custom written by the author. Although the tools were initially restricted to one software package, they were extended throughout the project to utilize the more versatile functionality of other specialized progammable software, which aided data management as well as data visualisation and statistical analysis.

The following sections provide an overview of the programming software and techniques used and outline their application in terms of field data processing and data preparation for analysis purposes. This is followed by an example which illustrates how the various software tools were integrated in order to complete the analysis tasks.

### 3.2.1 General description

Software

Three different software packages were used to develop a range of data management, visualisation and analysis tools. They are listed in Table 3-4, together with a short description of their main areas of application. VBA, IDL and EViews are all scriptable packages, i.e. they allow the automation of repetitive tasks, which make them valuable tools when dealing with large data sets. Note, routines written in VBA for Excel are commonly referred to as 'macros'.

Due to its spreadsheet and VBA capabilities, Excel provided a convenient platform for combining the collected raw data for each data run and for cataloguing the files. VBA macros aided the automation of repetitive tasks, such as downloading, converting and formatting of data after each data collection session. However, Excel proved to be impractical for advanced analytical procedures involving large data sets.

The progammable software IDL was found to be more suitable than Excel for the processing and analysis of large data sets. It also allowed the development of novel visualisation techniques which aided the exploratory analysis of the extensive and complex field study data set. Initially, however, Excel macros were used to extract relevant sub-sets of data as required for the analysis and to save them out in standard text format, ready to be accessed by routines from the other software packages. Later on, with improving programming skills, special IDL routines were written which accessed the raw data files directly and saved the extracted data as IDL array variables

which could be recalled later in subsequent analysis steps. Apart from using IDL for data extraction and to visualize data in different ways, it was also used to carry out most of the statistical analysis. This was done by combining basic statistical functions included in IDL to calculate complex statistical equations. An overview of the IDL routines written for analysis purposes is given in Section 3.2.3.

| Software package | Areas of application |
|---|---|
| **VBA**<br>Visual Basic for Applications, Microsoft® Excel, Microsoft Corporation | initial data manipulation and organisation<br>extraction of specific data for processing in IDL<br>running test cases for IDL statistics routines |
| **IDL**<br>Interactive Data Language from Research Systems, Inc. | visualisation<br>extraction of specific data for analysis<br>exploratory analysis<br>statistical analysis |
| **EViews**<br>Statistical Analysis Software, Quantitative Micro Software | additional statistical tests<br>running test cases for IDL statistics routines |

**Table 3-4 Programmable software used to develop specialized software tools**

Since the programming effort for more advanced statistical tests is relatively high in IDL, it was decided to use the statistical software package EViews where appropriate. For this purpose, code was added to the IDL routines in order to export the relevant data as text files, formatted to be read directly by custom written EViews routines. Results from the EViews calculations were then automatically saved out as text files again.

Final results from the analysis were generally saved out and stored as text files (tables) or images (charts) for future reference.

Quality assurance and methods

Automated data processing, using custom written programming code, helped to significantly reduce processing times and guaranteed that each collected data set was processed in exactly the same way. Thus the occurrence of accidental and unnoticed processing errors was minimized. In order to confirm that the code produced the desired results, each routine was subjected to a rigorous testing process before it was first applied. Throughout the development process, the routines were continually tested by inspecting the values of process variables. Further tests were carried out for

the completed routines. For data manipulation, these consisted of a visual inspection and comparison of data tables, while results from custom written analysis routines were cross checked with test cases in at least one of the other software packages. For example, Excel and EViews were used to confirm calculation results from IDL routines.

Following good programming practice, a number of 'reusable' routines were written which automated commonly executed tasks, such as reading data from a file, calculating descriptive statistics or printing charts. Since these could be called from other routines by using a minimum of additional code and testing, the development process could be simplified and accelerated.

For the same reasons, each routine was supported by documentation, usually included in the code as comments. For the most complex and the most frequently used processes, i.e. involving routines from different software packages, additional reference charts were compiled to enable efficient usage.

Most of the different analysis methods required that calculations were repeated for a number of different cases, often 'looping' through numerous sub-sets of data, i.e. consecutively selecting data for different conditions such as winter/summer, am/pm etc. Using a nested loop approach, with one or more 'central' visualisation or calculation sub-routines, ensured that results and chart files could automatically be generated for these data sets in an efficient and consistent manner. This was particularly advantageous in cases where modifications needed to be applied to the visualisation or calculation routines, since charts and results could easily be updated.

Due to the large number of files generated, it was necessary to adopt a consistent naming and storing system. This included naming output files in a way that provided information on how the files were generated, and what data was used. For example, a file named 'data_Upp_AM_ac_1.txt' would contain all raw data collected during morning runs, in anti-clockwise direction, on link 1 (Uppingham Road circuit). By naming all data and result files in a similar manner and placing them in appropriately named folders, it was ensured that they were easily accessible. Moreover, the file names could be used within the custom written routines to, for instance, derive title strings for charts and to automatically name result files and place them in specific folders. Although this may seem trivial, solving this naming issue was highly important for the efficient implementation of the analysis routines since files saved by

one routine could easily be read by other routines, using file names and locations as identifiers.

Throughout the project, some 50 main routines were written for data processing and analysis, each typically accompanied by various sub-routines. The main programming effort was focussed on developing analysis tools in IDL, which were supported by various EViews and VBA routines. Although the use of VBA macros declined throughout the project, due to improving IDL programming skills and advanced requirements, they played a vital role in the processing of data in the field study phase. The following section describes in detail how macros were used to aid the formatting and organizing of field data on a day to day basis.

### 3.2.2 Processing and organisation of the field data

The field study data consisted of eight data channels, four from the OSIRIS monitor (TSP[9], $PM_{10}$, $PM_{2.5}$ and $PM_1$), and one each from the P-Trak monitor, CO sensor and speed sensor.[10] It was considered desirable to combine these data channels into one data set (i.e. one for each data collection session), in order to reduce the number of stored files and to aid the subsequent analysis, which was likely to make use of two or more data channels at any one time. Due to the high logging frequency of 1 Hz, the amount of data collected was immense, so that it was not practical to store the raw data in a spreadsheet format. It was therefore decided to hold the data from each data collection session in separate text files, and to use a spreadsheet index to facilitate the selection of relevant data files for analysis. Keeping raw data in standard text file format ensured that it could be accessed by other applications as necessary and that file sizes were kept comparatively small.

In order to reduce data processing time and to minimize the chance of operator errors occurring, a semi-automated procedure was developed which utilized VBA macros in Excel. The macros were custom written to access the raw data from the individual monitoring devices and to organize it according to requirements (i.e. indexed text file collection). Figure 3-3 shows a simplified schematic of the steps involved in the data processing procedure, illustrating the use of macros and manual processing. As the

9. Total suspended particulates, i.e. all particles detected by the device.

10. Note, although TSP, $PM_1$ and CO were disregarded for analytical purposes, they were kept together with the other data channels throughout the data processing phase.

figure shows, the final field database consisted of three main parts: (1) the individual raw data files; (2) the index file; and (3) the event data files, which were generated as follows.

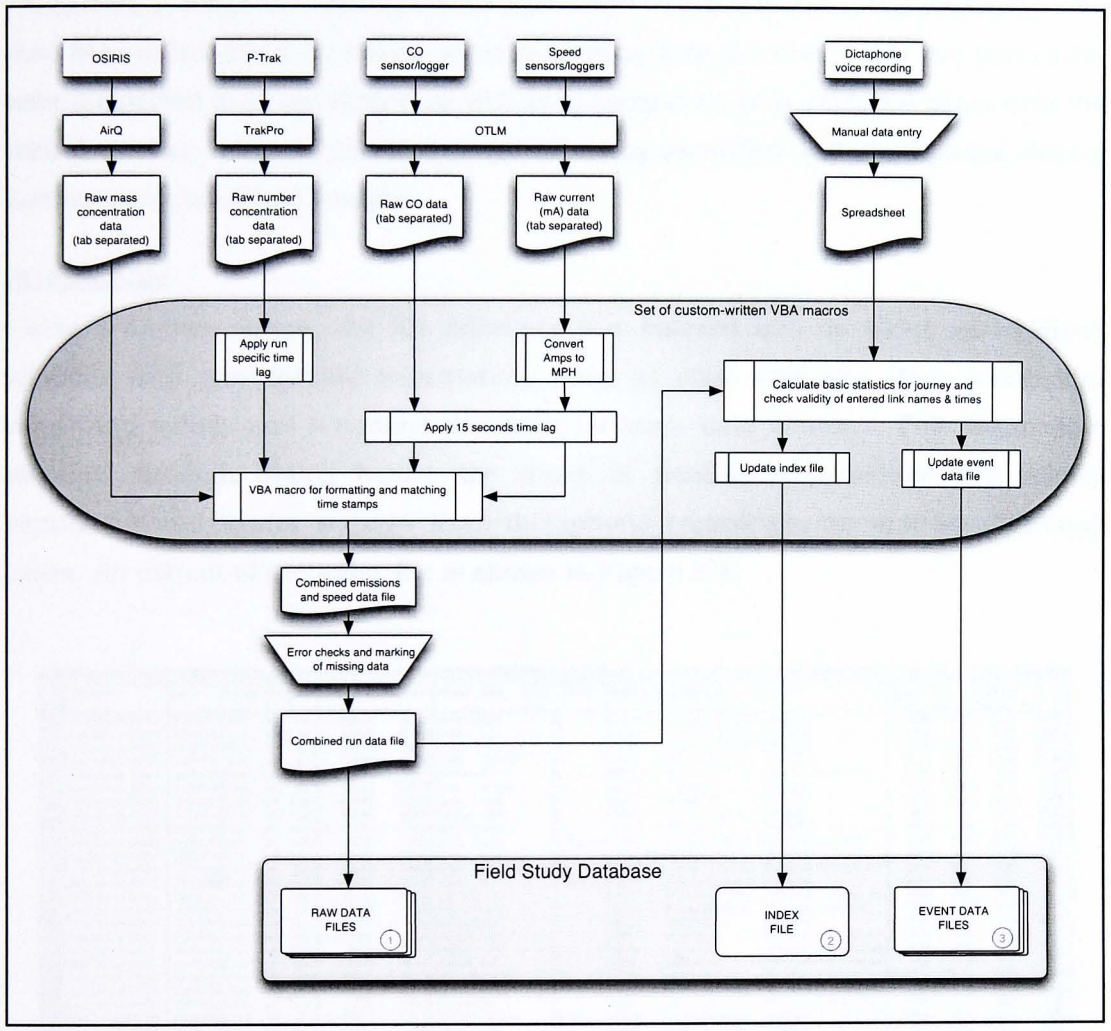


**Figure 3-3 Simplified schematic of field data processing procedures**

(1) Raw data files

Using the proprietary software for each monitor or data logger, the data files were converted into ASCII[11] text files. With custom written VBA macros these files were then reformatted and modified, correcting for known time lags of the data loggers, where applicable (i.e. the CO data was found to be constantly lagging by 15 seconds

11. American Standard Code for Information Interchange, standard plain text format.

due to a start-up delay). Time stamps in all four files were matched to compile a single file containing a complete set of data channels for the run.

Data fields in the combined file include time stamps, four OSIRIS data columns (TSP, $PM_{10}$, $PM_{2.5}$, $PM_1$), CO, instantaneous speed and P-Trak data (ultrafine particles). The data file was screened for errors, such as missing data (Section 3.1.6). All data which were suspected to be unreliable, as well as missing data, were replaced in the data file with a marker value so that they could be easily identified and disregarded during further processing and analysis.

(2) Index file

Using a further macro, the file reference was entered into an index spreadsheet together with run specific information, such as start and stop time, mean and maximum values and standard deviation for each data channel. The macro also allowed, through dialog boxes, the input of weather information and vehicle ventilation and heater settings from dictaphone recordings, as well as additional notes. An extract of the index file is shown in Figure 3-4.

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | file name | date | in/out | journ | vehi | start time | stop time | clouds | wind | temp | other | windows | ventilation | other | mean | in | ug/m^3 | |
| 2 | | | | | | | | | | C | | | | | TPM | PM10 | PM2.5 | PM1 |
| 9 | town_pm van 070502 | 07/05/2002 | town | pm | van | 17:04:41 | 18:21:51 | mixed clou | light | 15 | | closed | mech vent +1, heater of | PM1 count | 102.58 | 53.91 | 23.18 | 25.05 |
| 10 | town_pm van 080502 | 08/05/2002 | town | pm | van | 16:51:18 | 18:02:54 | cloudy | light | 18 | | closed | mech vent +1, heater | PM1 count | 95.64 | 51.23 | 17.33 | 19.28 |
| 11 | town_pm van 090502 | 09/05/2002 | town | pm | van | 16:48:42 | 18:01:06 | cloudy | no wind | 14 | | closed | mech vent +1, heater | PM1 count | 243.17 | 180.79 | 112.34 | 90.55 |
| 12 | town_am van 100502 | 10/05/2002 | town | am | van | 7:35:16 | 9:10:12 | cloudy | no wind | 12 | | closed | mech vent +1, heater | PM1 count | 224.06 | 146.14 | 87.96 | 86.24 |
| 13 | town_pm van 150502 | 15/05/2002 | town | pm | van | 17:05:30 | 18:14:01 | mixed clou | ? | 22 | | half open | mech vent +1 | PM1 count, no Ptrak data | 111.36 | 64.22 | 13.61 | 7.81 |
| 14 | town_am van 160502 | 16/05/2002 | town | am | van | 7:52:04 | 9:00:55 | sunny | no wind | 13 | | closed | mech vent +1 | PM1 count | 145.62 | 64.84 | 16.79 | 12.13 |
| 15 | town_am van 160502 | 16/05/2002 | town | am | van | 16:48:47 | 17:52:12 | sunny | windy | 25 | | open | mech vent +1 | PM1 count | 100.63 | 56.67 | 12.45 | 9.88 |
| 16 | town_am van 170502 | 17/05/2002 | town | am | van | 7:50:45 | 8:59:34 | overcast | windy | 16 | | closed | mech vent +1, heater | PM1 count | 84.61 | 45.51 | 13.67 | 10.79 |
| 17 | town_am van 270502 | 27/05/2002 | town | am | van | 17:06:00 | 18:17:30 | mixed clou | windy | 20 | | open | mech vent +1 | PM1 count | 81.66 | 42.18 | 7.85 | 5.45 |
| 18 | town_am van 290502 | 29/05/2002 | town | am | van | 7:51:18 | 9:05:09 | cloudy | windy | 13 | showers | closed | mech vent +1, heater | PM1 count, no CO data | 48.86 | 30.35 | 10.68 | 7.33 |
| 19 | town_pm van 290502 | 29/05/2002 | town | pm | van | 16:49:55 | 18:09:10 | sunny | windy | 18 | | open/close | mech vent +1 | PM1 count | 78.53 | 41.13 | 10.10 | 5.94 |
| 20 | town_am van 310502 | 31/05/2002 | town | am | van | 7:50:06 | 8:52:35 | sunny | light wind | 12 | | closed | mech vent +1, heater | PM1 count, VERY quiet BH? | 85.79 | 52.76 | 19.55 | 12.39 |
| 21 | town_am van 100602 | 10/06/2002 | town | am | van | 16:50:22 | 18:19:09 | cloudy | ? | 14 | rainy | closed | mech vent +1, heater | PM1 count | 83.69 | 43.56 | 12.80 | 8.25 |
| 22 | town_am van 120602 | 12/06/2002 | town | am | van | 7:57:36 | 8:50:48 | mixed clou | windy | ? | | closed | mech vent +1, heater | PM1 count, no CO data, foot | 90.23 | 49.03 | 13.01 | 8.65 |
| 23 | town_am van 120602 | 12/06/2002 | town | pm | van | 16:54:48 | 18:05:19 | overcast | windy | 21 | | closed | mech vent +1, heater | PM1 count, no CO data | 79.61 | 41.33 | 10.42 | 6.38 |
| 24 | town_am van 140602 | 14/06/2002 | town | am | van | 7:53:27 | 8:59:17 | overcast | light | 18 | | closed | mech vent +1 | PM1 count | 92.84 | 54.20 | 18.15 | 15.64 |
| 25 | town_pm van 170602 | 17/06/2002 | town | pm | van | 16:53:29 | 18:13:54 | mixed clou | windy | 24 | | open | mech vent +1 | PM1 count, building works Al | 61.39 | 31.41 | 7.45 | 6.56 |
| 26 | town_am van 190602 | 19/06/2002 | town | am | van | 8:02:39 | 9:10:48 | mixed sun | windy | 19 | | closed | mech vent +1 | PM1 count, only UppRd data | 97.92 | 52.82 | 16.27 | 9.25 |
| 27 | town_pm van 190602 | 19/06/2002 | town | pm | van | 16:59:07 | 17:59:37 | overcast | light | 24 | | closed | mech vent +1 | PM1 count, only UppRd data, | 75.13 | 38.41 | 9.67 | 5.97 |
| 28 | town_am van 250602 | 25/06/2002 | town | am | van | 8:10:54 | 9:03:24 | mixed clou | no wind | 16 | | closed | mech vent +1 | PM1 count, no AbbLane data | 120.84 | 69.48 | 19.74 | 12.80 |
| 29 | town_am van 260602 | 26/06/2002 | town | am | van | 8:04:52 | 8:56:42 | sunny | light | ? | | closed | mech vent +1, heater | PM1 count, no AbbLane data | 63.55 | 35.58 | 10.32 | 8.36 |
| 30 | town_pm van 260602 | 26/06/2002 | town | pm | van | 17:01:16 | 17:58:51 | mixed clou | light | 22 | | open | mech vent +1 | PM1 count, no AbbLane data | 127.66 | 67.37 | 13.18 | 8.11 |
| 31 | town_am van 030702 | 03/07/2002 | town | am | van | 7:50:24 | 9:05:05 | cloudy | no wind | 14 | rainy | closed | mech vent +1, heater | PM1 count | 111.79 | 65.90 | 17.44 | 12.86 |
| 32 | town_pm van 030702 | 03/07/2002 | town | pm | van | 17:03:44 | 18:22:34 | raining | light | 15 | | closed | mech vent +1, heater | PM1 count | 74.61 | 41.88 | 11.14 | 7.96 |
| 33 | town_am van 040702 | 04/07/2002 | town | am | van | 7:50:47 | 9:02:33 | sunny | light | 15 | | closed | mech vent +1, heater | PM1 count | 79.80 | 41.95 | 10.97 | 6.84 |
| 34 | town_am van 040702 | 04/07/2002 | town | pm | van | 17:04:55 | 18:11:40 | overcast | light | 19 | | closed | mech vent +1 | PM1 count | 78.18 | 41.52 | 9.96 | 7.02 |
| 35 | town_am van 100702 | 10/07/2002 | town | am | van | 8:02:20 | 9:18:14 | sunny | no wind | 15 | | closed | mech vent +1, heater | PM1 count | 116.88 | 60.15 | 11.82 | 8.56 |
| 36 | town_pm van 110702 | 11/07/2002 | town | pm | van | 16:53:24 | 18:06:00 | overcast | light | 19 | | closed | mech vent +1 | PM1 count | 104.34 | 45.22 | 8.87 | 6.09 |
| 37 | town_am van 170702 | 17/07/2002 | town | am | van | 7:51:35 | 8:55:04 | overcast | light | 16 | | closed | mech vent +1, heater | PM1 count, school holidays | 99.62 | 54.50 | 14.78 | 11.66 |
| 38 | town_pm van 170702 | 17/07/2002 | town | pm | van | 16:57:02 | 18:08:35 | overcast | light | 24 | | closed | mech vent +1 | PM1 count, school holidays | 83.82 | 51.56 | 19.02 | 18.60 |
| 39 | town_pm van 180702 | 18/07/2002 | town | pm | van | 16:57:37 | 18:11:47 | overcast | light | 21 | | half oven | mech vent +1 | PM1 count, school holidays | 129.08 | 67.98 | 13.70 | 11.81 |
| 40 | town_am van 190702 | 19/07/2002 | town | am | van | 8:03:43 | 9:08:08 | overcast | light | 17 | | closed | mech vent +2 | PM1 count, school holidays | 103.97 | 54.84 | 12.27 | 8.21 |
| 41 | town_am van 240702 | 24/07/2002 | town | am | van | 7:49:17 | 9:01:36 | overcast | light | 17 | | closed | mech vent +1, heater | PM1 count, school holidays | 107.84 | 63.71 | 18.22 | 10.78 |
| 42 | town_pm van 240702 | 24/07/2002 | town | pm | van | 16:53:15 | 18:04:09 | mixed sun | windy | 23 | | closed | mech vent +1 | PM1 count, school holidays | 54.16 | 28.85 | 7.58 | 5.28 |
| 43 | town_am van 250702 | 25/07/2002 | town | am | van | 7:53:19 | 8:57:07 | sunny | no wind | 15 | | closed | mech vent +2 | PM1 count, school holidays | 109.58 | 54.18 | 12.57 | 9.30 |
| 44 | town_pm van 250702 | 25/07/2002 | town | pm | van | 16:56:28 | 18:04:07 | overcast | windy | 22 | | half open | mech vent +1 | PM1 count, school holidays | 75.12 | 40.92 | 11.91 | 10.89 |
| 45 | town_am van 010802 | 01/08/2002 | town | am | van | 7:53:14 | 8:56:45 | overcast | no wind | 19 | occasion | closed | mech vent +1, heater | PM1 count, school holidays | 133.84 | 81.94 | 22.18 | 21.36 |
| 46 | town_pm van 010802 | 01/08/2002 | town | pm | van | 16:47:22 | 18:02:17 | mixed sun | no wind | 21 | | closed | mech vent +1 | PM1 count, school holidays | 97.59 | 49.64 | 12.08 | 9.33 |
| 47 | town_am van 020802 | 02/08/2002 | town | am | van | 8:06:18 | 9:11:01 | sunny | no wind | 16 | | closed | mech vent +1, heater | PM1 count, school holidays | 161.95 | 83.15 | 20.09 | 17.80 |
| 48 | town_am van 070802 | 07/08/2002 | town | am | van | 7:56:30 | 9:05:06 | sunny | no wind | 20 | | closed | mech vent +1 | PM1 count, school holidays | 182.43 | 126.57 | 67.79 | 78.86 |
| 49 | town_am van 070802 | 07/08/2002 | town | am | van | 16:49:54 | 18:16:40 | overcast | windy | 27 | showery r | closed | mech vent +1 | PM1 count, school holidays, | 91.96 | 53.98 | 16.69 | 15.12 |
| 50 | town_pm van 080802 | 08/08/2002 | town | pm | van | 16:53:18 | 18:06:43 | overcast | light | 20 | | open/close | mech vent +1 | PM1 count, school holidays | 119.75 | 64.38 | 13.20 | 10.16 |

**Figure 3-4 Extract from index file**

This spreadsheet acts as an index to the database of raw data files. It can be filtered using certain criteria, such as weather conditions, high or low values etc., in order to,

for example, select data files that were recorded under similar conditions or contain similar values.

(3) Event data file

For each run, event data, which had been recorded onto a dictaphone during data collection, was entered manually into an Excel spreadsheet. Using two macros, the data were formatted and appended on to a route specific event data file (i.e. Upp.xls, Abb.xls, Cit.xls) which contained all event data for runs carried out on this particular route. Figure 3-5 shows an example of event data as stored in the spreadsheet.

| | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | | | | | | | | | |
| 2 | link number | date and time | time | time on link | average speed | notes | | | | |
| 2499 | | | | | | | | | | |
| 2500 | start time | 13/11/02 16:56:20 | 16:56:20 | | | | | | | |
| 2501 | gap | 13/11/02 16:56:20 | 16:56:20 | 00:00:22 | -9999.00 | G regC | | | | |
| 2502 | 21 | 13/11/02 16:56:42 | 16:56:42 | 00:00:20 | 47.21 | nt | | | | |
| 2503 | 22 | 13/11/02 16:57:02 | 16:57:02 | 00:00:54 | 17.16 | | | | | |
| 2504 | 23 | 13/11/02 16:57:56 | 16:57:56 | 00:00:25 | 45.95 | N regC | | | | |
| 2505 | 24 | 13/11/02 16:58:21 | 16:58:21 | 00:01:29 | 27.93 | F regC 2.43 | N regC 3.10 | | | |
| 2506 | 25 | 13/11/02 16:59:50 | 16:59:50 | 00:00:16 | 44.77 | nt | | | | |
| 2507 | 26 | 13/11/02 17:00:06 | 17:00:06 | 00:00:28 | 50.68 | | | | | |
| 2508 | 27 | 13/11/02 17:00:34 | 17:00:34 | 00:00:33 | 63.36 | R regC | | | | |
| 2509 | 28 | 13/11/02 17:01:07 | 17:01:07 | 00:00:15 | 32.05 | nt | | | | |
| 2510 | 29 | 13/11/02 17:01:22 | 17:01:22 | 00:00:22 | 36.01 | M reg cab 5.10 | | | | |
| 2511 | 30 | 13/11/02 17:01:44 | 17:01:44 | 00:00:07 | 45.64 | G regC | | | | |
| 2512 | 31 | 13/11/02 17:01:51 | 17:01:51 | 00:00:42 | 14.95 | sb | | | | |
| 2513 | 21 | 13/11/02 17:02:33 | 17:02:33 | 00:00:37 | 25.52 | sb | | | | |
| 2514 | 22 | 13/11/02 17:03:10 | 17:03:10 | 00:00:46 | 20.15 | sb | J regC 7.04 | Q back at lights | | |
| 2515 | 23 | 13/11/02 17:03:56 | 17:03:56 | 00:00:45 | 25.53 | sb | | | | |
| 2516 | 24 | 13/11/02 17:04:41 | 17:04:41 | 00:03:14 | 12.81 | sb | Q back | K regC 9.55 | Q back at lights | |
| 2517 | 25 | 13/11/02 17:07:55 | 17:07:55 | 00:00:16 | 44.77 | nt | | | | |
| 2518 | 26 | 13/11/02 17:08:11 | 17:08:11 | 00:01:12 | 19.71 | N regC | nt 12.36 | | | |
| 2519 | 27 | 13/11/02 17:09:23 | 17:09:23 | 00:00:34 | 61.50 | L regC | | | | |
| 2520 | 28 | 13/11/02 17:09:57 | 17:09:57 | 00:00:48 | 10.02 | P regC | | | | |
| 2521 | 29 | 13/11/02 17:10:45 | 17:10:45 | 00:00:17 | 46.60 | sb | | | | |
| 2522 | 30 | 13/11/02 17:11:02 | 17:11:02 | 00:00:09 | 35.50 | sb | | | | |
| 2523 | 31 | 13/11/02 17:11:11 | 17:11:11 | 00:01:37 | 6.48 | sb | Q back | | | |
| 2524 | 21 | 13/11/02 17:12:48 | 17:12:48 | 00:00:24 | 39.34 | P regC | H regC 16.42 | | | |
| 2525 | end time | 13/11/02 17:13:12 | 17:13:12 | | | | | | | |
| 2526 | | | | | | | | | | |
| 2527 | | | | | | | | | | |

raw data / links AM / links PM / met data / collated link data / journey data for IDL / raw link data for IDL / raw run data for IDL PM10

**Figure 3-5 Example of event data record**

Additionally, the two event data macros also calculated the average link speed and how much time was spent on each link, and included the results in the event data record. The macros also served as a means of verifying the entered time values by generating error messages in cases where typing or recording errors resulted in non-sensible time values. If errors were detected, they were corrected manually and the macros were run again.

After the completion of the field study campaign, the final versions of the event data files were exported as text files in order to be accessible to analysis routines.

Accessing field data

Due to its design, the field study database allows programming routines to easily access the data set in order to select raw data for use in the analysis. For example, to select raw data for a particular link on a particular run, the data loading routine would read start and end times for each link from the event data file. Based on the date and time, the appropriate raw data file would be loaded (facilitated by naming convention, e.g. town_am van 081002.txt). By matching the start and end times from the event data file to time stamps in the raw data file, the pollution and speed data collected on the required link could be extracted.

During the analysis phase this process was frequently employed, typically using IDL routines to automatically loop through multiple links and runs. A detailed description of the use of custom written routines for analysis purposes in given in the next section.

### 3.2.3 Programming routines for the data analysis

As noted above, data analysis was carried out using mainly custom written IDL routines. Some 30 main routines have been written, each typically accompanied by various sub-routines. Figures 3-6 and 3-7 give an overview of the routines written for the various data analysis methods used, the former corresponding to the analysis presented in Chapters 4 and the latter to the methods employed in Chapters 5 and 6.

The figures show the main routines in dark gray and, where applicable, associated sub-routines in light gray. The routines are grouped in boxes indicating the type of analysis they were used for. These boxes thus loosely correspond to the headings in the analysis chapters of this thesis. Figure 3-8 lists further utility routines which were used to execute basic recurring tasks such as, for example, extracting data from the raw data files or plotting a certain type of chart. These 'toolbox' functions are called by many of the other routines but are not included in Figures 3-6 and 3-7 to simplify the diagrams.

The names of the routines listed in the figures are the actual IDL procedure names, excluding the extension (*.pro), which were generally chosen to be descriptive in order to aid identification. Since this set of programming routines evolved with increasing programming skills and advanced analytical requirements, the names do not follow a consistent naming convention, apart from using the prefix 'func' to mark sub-functions, as opposed to the more complex sub-procedures.

**Figure 3-6 IDL routines for data analysis methods used in 'Overview and General Data Analysis' (Chapter 4)**

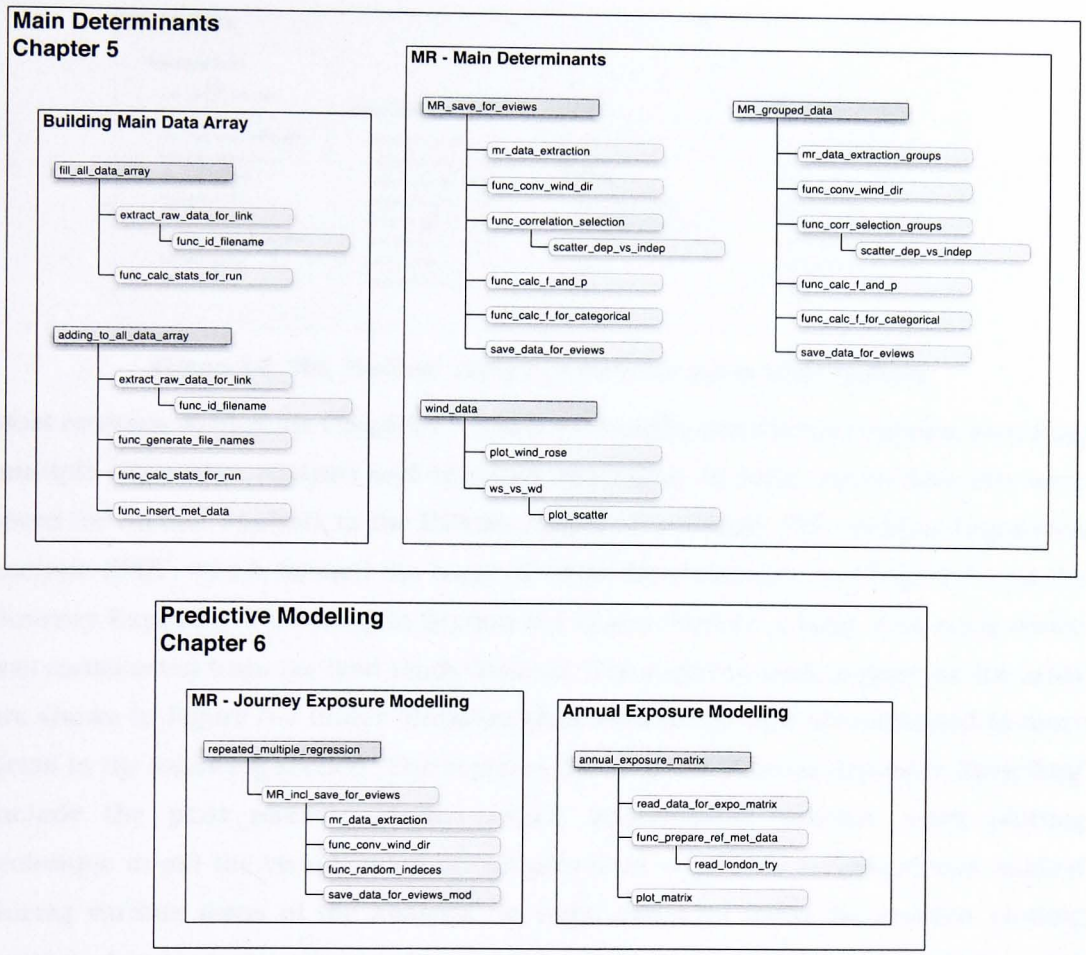


**Figure 3-7 IDL routines for data analysis methods used in 'Identifying the Main Determinants' (Chapter 5) and 'Predictive Modelling' (Chapter 6)**

The routines aided the execution of four general types of tasks: data extraction; data manipulation; calculations; and output of results. The latter involved printing the result tables to text files or plotting charts or a combination of both. As the figures illustrate, typically one main routine was written which proceeded through all the tasks required for a certain analysis purpose, often calling specialised sub-routines for data extraction, calculation and/or output of results.

The routines written for Chapter 4 included a wide range of calculation and plotting procedures (Figure 3-6). While 'Descriptive Statistic' and 'Correlation Analysis' required mainly calculation routines, other task groups included complex plotting routines. For instance, novel data visualisation techniques were implemented with the plotting routines included in the boxes 'Time Series Plots' and 'Stacked Histograms'. Examples of outputs from those routines are shown in Sections 4.3 and 4.6.

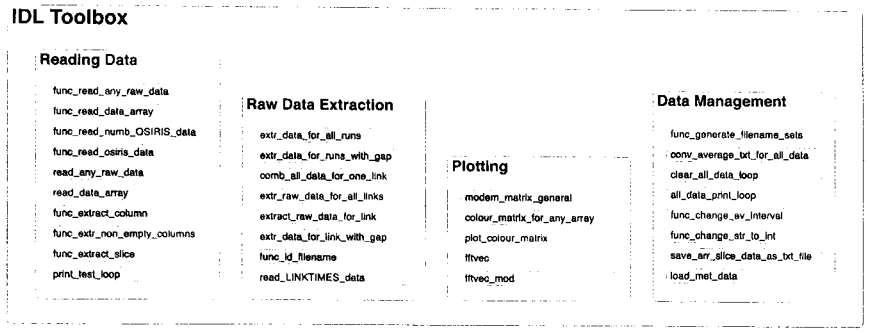


**Figure 3-8 'IDL Toolbox' utility functions for use in other routines**

Most routines written for Chapters 5 and 6 are lengthy calculation routines, executing multiple regression analysis and exposure modelling. In some cases, data sets were saved for further analysis in the EViews statistical package. The multiple regression analysis ('MR') which formed the basis of 'Main Determinants' in Chapter 5 and the 'Journey Exposure Modelling' in Section 6.1 both relied on a large data array which was constructed from the field study data set. The routines used to generate the array are shown in Figure 3-7 under 'Building Main Data Array' and are explained in more detail in the following section. The routines listed under 'Annual Exposure Modelling' include the `plot_matrix` routine which implemented another novel plotting technique to aid the visualisation of complex data sets. This technique was applied during various steps of the analysis, in most cases by using the toolbox plotting routine `plot_colour_matrix`. For example, Section 4.4 contains an output from the routine for the time-lag analysis results.

The figures listing the custom written routines, and the brief description of their function given in this section, are intended to provide a general overview of the custom written tools used for data processing and the analysis. Although the development of the programmed tools was a lengthy process, which required an intense learning effort, it was greatly beneficial. It allowed the exploration of many analysis approaches which required time consuming, and thus virtually impracticable, data selection processes and calculations. The following section describes in more detail one of the sets of routines, focussing on the construction of the main data array which formed the basis of the multiple regression analysis.

### 3.2.4 Example: Construction of the main data array

The routines described here were written in order to generate a data array that contained all relevant data for the multiple regression analyses (Chapter 5). The aim was to store descriptive statistics (e.g. mean, standard deviation etc.) of the pollution data collected for each link on each data collection run. The array was further to include link specific parameters (e.g. link length, road function etc.) and run specific parameters (e.g. date and meteorological data etc.).

Various other options for storing the required data in an easily accessible format had been considered in preparation for the analysis. Concepts explored included using individual files with meaningful file names, as discussed above, or creating small arrays for each run in combination with header files (storing string information about run parameters). However, it was decided that one large array which would contain all relevant link and run specific statistics would be most suitable for the use by programming routines. This approach allowed the implementation of efficient programming loops for the selection of data from multiple links and runs by using simply to define integer loop variables (easier to implement than if using strings) and did not require the opening of multiple text files during each analysis session.

The main challenge was to construct the array so that all link specific data, as well as run specific date parameters and measured pollution and meteorological data, could be stored in a logical way. It was vital that the array was easy to interpret in order to aid the writing of subsequent routines that accessed the array. It was recognized that this may lead to a very regular structure with relatively high redundancy. However, since the ease of processing was of paramount importance, the overhead of 'empty' sections of the array was considered trivial. Moreover, setting up an array slightly larger than initially required aided the inclusion of further parameters at a later stage if found necessary.

All these considerations led to the construction of a 3-dimensional array as shown in Figure 3-9. Two tables have been included in the figure to provide details of the location of the various parameters in the array. The first dimension of the array consists of two parts: the left part (dim1 = 0 to 49) for storing link specific information and the right part (dim1 = 50 to 249) for run data[12]. The second

12. The index subscripts in IDL start with 0 for the first element in the array.

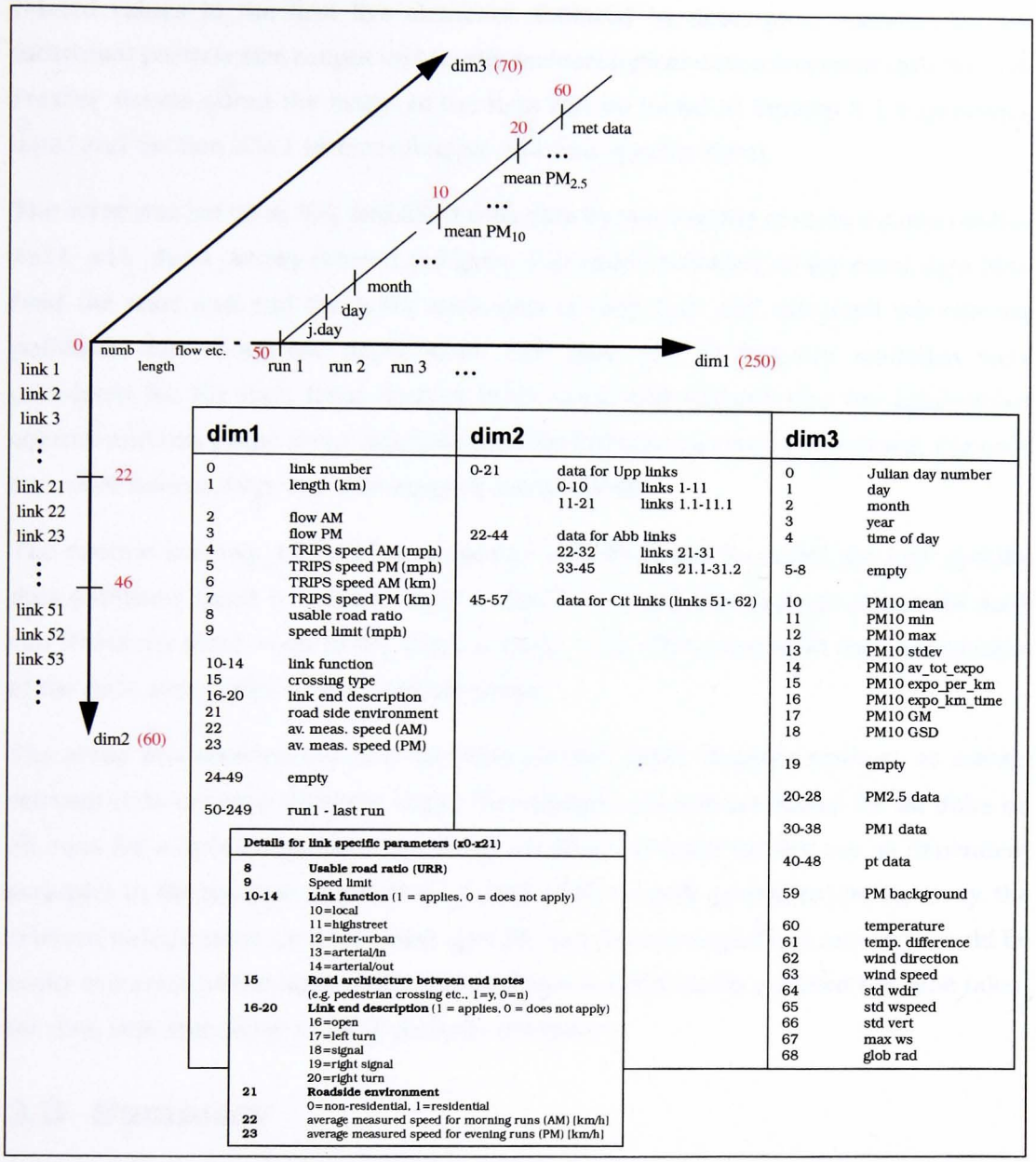


**Figure 3-9 Main IDL data array**

dimension contains the individual links in ascending order. Since clockwise and anti-clockwise collected data was to be treated separately, each link was included twice, once as 'x' (anti-clockwise) and once, as explained above, with an additional marker as 'x.1' or 'x.2' (clockwise). The second dimension therefore starts with anti-clockwise links on Route 1 followed by the clockwise runs on the same route, and so on for the other routes. As there are 58 links, two empty rows remain at the bottom of the array. The third dimension holds detailed information for each run, starting with date

related values in the first five elements, followed by descriptive statistics for the individual particle size ranges and finally meteorological data associated with the run. Further details about the origin of the data can be found in Section 3.1.5 (pollution data) and Section 5.3.1 (meteorological and link specific data).

The array was set up in IDL and filled with data by running the custom written routine **fill_all_data_array** (shown in Figure 3-7) which accessed all the event data files, read the start and end times for each visit of each link and extracted the relevant pollutant data from the appropriate raw data file. Descriptive statistics were calculated for the data from each of these visits and entered into the appropriate column and row of the array, depending on the link and run being processed. For each run, date information was also entered automatically.

The routine **adding_to_all_data_array** was then run to insert the link specific data (columns dim2 = 0 to 24, dim1 = dim3 = 0) and meteorological data for each run (columns dim3 = 60 to 68, dim1 = dim2 = 0). This concluded the construction of the data array prior to the analysis phase.

The array was constructed in a way that allowed other analysis routines to extract relevant data in many different ways. For example, all average values for all links on all runs for a specific particle size range could be selected for the use as dependent variables in the multiple regression analysis. Due to their placement in the array, the relevant independent variables (link specific and meteorological parameters) could be easily extracted at the same time. This approach significantly reduced the time taken for data selection prior to each analysis procedure.

## 3.3 Summary

A field study methodology was developed and successfully applied during an extensive field study campaign. Three field study routes, split into individual road links, were devised. Over the period of one year, the routes were visited regularly during morning and evening rush hour. At each visit, concentrations of particulate matter were measured and related event data was recorded.

The field campaign generated an extensive data set. Custom written data processing routines were used to process the data and to store it the field study data base. Further programming tools were developed which aided data extraction,

manipulation and calculations during the analysis phase as well as the output of results in chart and table format.

The analysis methods used and the results and charts derived with the aid of these custom written routines are explained in detail in the following chapters.

Chapter

# 4 Overview and General Data Analysis

A comprehensive data set of driver exposure to particulate matter was acquired during a twelve month field study (Chapter 3). Due to the high logging frequency used for the measurements, the data contain more detail regarding the temporal dynamics of in-vehicle concentrations of particulate matter than the data from most other exposure studies (Section 2.3.3). Moreover, since measurements were taken for three particle size fractions (ultrafine, $PM_{10}$, $PM_{2.5}$), the data set can be used to explore whether any observed exposure patterns are consistent for all particle sizes.

Based on the event data regarding the road links on the collection routes, all acquired personal exposure data can be associated with the road section it was collected on. This aids the investigation of the effect of road specific parameters on exposure values. Since the field campaign extended over a whole year, data was collected under a wide range of meteorological conditions. The data set can therefore also be used to study the influence of meteorological parameters on driver exposure.

In order therefore to best take advantage of this extensive data set, the following objectives were set for the analysis:

(1) Examine the field data in its raw form, as high resolution time series, in order to gain an understanding of the short-term behaviour of in-vehicle particle concentrations.

(2) To investigate specifically selected sub data sets and processed metrics, such as journey averages, with the aim to identify patterns which may indicate potential underlying relationships between in-vehicle concentrations and external factors.

(3) To use multiple regression analysis to investigate whether the observations from (1) and (2) can be confirmed and quantified by identifying the main determinants of driver exposure.

The first two objectives are the focus of this chapter while the methods and results regarding the third are explained in Chapter 5.

## 4.1 Introduction

In order to investigate the dynamics of driver exposure, a stratified approach to analysing the field data was formulated. This is described below.

Descriptive statistics were calculated for sub-sets of the data in order to derive an overview of the seasonal and diurnal variations of driver exposure to particulate matter. In order to uncover underlying patterns in the data set and to aid the selection of appropriate analysis techniques, various visualisation techniques were employed to inspect the field data, from simple time series plots, showing every measured data point, to aggregated plots of histograms for grouped data.

Data from individual runs were used in a time series analysis, aiming to investigate underlying patterns on a high resolution short-term basis, such as the relationships between the different particle size ranges and vehicle speed. A method was developed that allowed the detection of peaks in a large number of time series measurements of concentrations. This facilitated the investigation of the temporal dynamics of driver exposure to particulate matter.

The main focus of the analysis was the investigation of driver exposure to ultrafine particles (UF). However, where appropriate, data for larger particles were considered for comparative purposes. Although concentrations of four size ranges of particles were logged by the OSIRIS particle monitor, only mass concentrations of $PM_{10}$ and $PM_{2.5}$ were included in the data analysis. Data for the coarser size range of total suspended particulates (TSP), which includes all particles with a diameter up to 20 $\mu m$, were not used due to the focus of the study being on small particles. Since the

number concentrations of $PM_1$ measured with the OSIRIS are only based on particles with a diameter greater than 0.4 μm, they typically result in very low counts values and do not include ultrafine particulate matter, i.e. particles with a diameter less than 0.1 μm. The data therefore provided little additional detail, particularly considering that the P-Trak measurements provided number concentration data which included particles with diameters greater than 0.02 μm, and were thus excluded from the analysis.

## 4.2 General statistics

As outlined in Section 3.1.7, the final data set comprises in-vehicle concentration data from 133 data collection sessions, i.e. 308 runs. A 'run' typically consists of data from two consecutively completed iterations of a route during the same am or pm session.

In order to get a first overview of the collected data, descriptive statistics were calculated. To retain some of the detail inherent in the data, statistics were generated for sub-sets of grouped data. The parameters and categories chosen to select sub-sets of data are presented in Table 4-1. They were selected based on the following considerations. Grouping data by when they were collected, i.e. season and time of day, may allow an investigation of seasonal and diurnal variability in the data. Similarly, grouping data by where it was collected, i.e. route, may give a first indication of the influence of road related parameters on exposure values. The data were further grouped by the direction of travel during the data collection run in order to account for variations between clockwise and anti-clockwise links.

| Parameter | Categories | Abbreviation |
|---|---|---|
| season | summer | s |
| | winter | w |
| time of day | morning rush-hour | am |
| | evening rush-hour | pm |
| location | Route 1 (Uppingham Road) | R1 or Upp |
| | Route 2 (Abbey Lane) | R2 or Abb |
| | Route 3 (City Centre) | R3 or Cit |
| direction | anti-clockwise | acw |
| | clockwise | cw |

**Table 4-1 Categories for the selection of sub-sets of field data**

For each sub-set, all valid data were extracted from the raw data files using custom written IDL routines. Geometric mean and standard deviation values were then calculated for each set. The results are presented in Table 4-2.

Mean particle concentrations for grouped data range from about 46 $x10^3$ to 116 $x10^3$ $cm^{-3}$ for ultrafine particles and from 11.2 to 25.3 $\mu g\, m^{-3}$ for $PM_{2.5}$. Mean concentrations recorded for $PM_{10}$ were between 27.9 and 77.9 $\mu g\, m^{-3}$.

Generally, the average values for driver exposure to $PM_{10}$ and $PM_{2.5}$ observed during the field study are comparable to results from other studies. For example, Gulliver and Briggs (2004) measured in-car mean concentrations of 43.16 and 15.54 $\mu g\, m^{-3}$ for $PM_{10}$ and $PM_{2.5}$, respectively, which are well within the range of average values of the present study. The results for in-car concentrations of $PM_{2.5}$ published by Adams et al. (2001a) ranged from 23.7 to 35.0 $\mu g\, m^{-3}$ and were thus slightly higher than in the present study, which could be due to traffic emissions being higher in central London, where that study was carried out.

As far as is known, no published number concentration data for driver exposure to ultrafine particles were available for comparison up to the time of completion of the study (June 2004). However, Bukowiecki et al. (2003) report background particle concentrations in the centre of Zürich of approximately 35 $x10^3$ $cm^{-3}$ ($\pm$ 5 $x10^3 cm^{-3}$) and in-traffic concentrations of > 80 $x10^3$ $cm^{-3}$. Preliminary results from the DAPPLE project[1] indicate that in-vehicle concentrations measured during journeys along Marylebone Road in London were around 95 $x10^3$ $cm^{-3}$ and thus well within the range of the concentrations observed during the present study. Average outside concentrations (roof top) measured in Leipzig, Germany, by Wehner and Wiedensohler (2003) were clearly lower (between 13 $x10^3$ to 19 $x10^3$ $cm^{-3}$), however, they were not solely based on rush hour data but included data from 24 hour measurements (i.e. including off peak and night time periods).

1. Private communication Surbjit Kaur, Deputy Scientific Coordinator (DAPPLE Project), Imperial College London. First results are published in Arnold et al., 2004.

| | | summer am | | | | summer pm | | | | winter am | | | | winter pm | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $n_s$ | UF | $PM_{10}$ | $PM_{2.5}$ | $n_s$ | UF | $PM_{10}$ | $PM_{2.5}$ | $n_s$ | UF | $PM_{10}$ | $PM_{2.5}$ | $n_s$ | UF | $PM_{10}$ | $PM_{2.5}$ |
| | | [#] | [# $cm^{-3}$] | [$\mu g\ m^{-3}$] | [$\mu g\ m^{-3}$] | [#] | [# $cm^{-3}$] | [$\mu g\ m^{-3}$] | [$\mu g\ m^{-3}$] | [#] | [# $cm^{-3}$] | [$\mu g\ m^{-3}$] | [$\mu g\ m^{-3}$] | [#] | [# $cm^{-3}$] | [$\mu g\ m^{-3}$] | [$\mu g\ m^{-3}$] |
| **Route 1 Upp. Rd.** | **cw** | 16 | 74 x$10^3$ | 57.0 | 19.9 | 20 | 60 x$10^3$ | 43.9 | 14.2 | 8 | 93 x$10^3$ | 39.4 | 18.3 | 8 | 58 x$10^3$ | 28.7 | 12.9 |
| | | | (1.6) | (1.9) | (1.9) | | (1.7) | (2.0) | (2.0) | | (1.7) | (1.8) | (1.5) | | (1.6) | (2.0) | (1.8) |
| | **acw** | 15 | 51 x$10^3$ | 39.2 | 13.5 | 16 | 46 x$10^3$ | 33.6 | 9.7 | 8 | 76 x$10^3$ | 34.0 | 15.6 | 8 | 55 x$10^3$ | 27.9 | 13.6 |
| | | | (1.8) | (2.0) | (1.6) | | (1.6) | (2.1) | (1.6) | | (2.0) | (2.2) | (1.9) | | (1.8) | (2.2) | (1.8) |
| **Route 2 Abb. Lane** | **cw** | 15 | 88 x$10^3$ | 77.9 | 25.3 | 16 | 61 x$10^3$ | 49.4 | 14.7 | 13 | 116 x$10^3$ | 57.1 | 23.5 | 13 | 93 x$10^3$ | 51.1 | 21.7 |
| | | | (1.5) | (1.9) | (1.9) | | (1.6) | (1.9) | (1.5) | | (1.5) | (1.8) | (1.5) | | (1.5) | 2.0 | (1.8) |
| | **acw** | 16 | 80 x$10^3$ | 58.9 | 18.3 | 16 | 66 x$10^3$ | 46.8 | 13.4 | 13 | 101 x$10^3$ | 49.4 | 20.1 | 12 | 88 x$10^3$ | 47.1 | 19.1 |
| | | | (1.5) | (1.9) | (1.5) | | (1.7) | (2.0) | (1.8) | | (1.6) | (1.8) | (1.5) | | (1.6) | 1.9 | (1.7) |
| **Route 3 City Cent.** | **acw** | 33 | 63 x$10^3$ | 45.3 | 16.0 | 32 | 48 x$10^3$ | 35.8 | 11.2 | 16 | 81 x$10^3$ | 43.1 | 19.6 | 14 | 71 x$10^3$ | 42.0 | 19.0 |
| | | | (1.7) | (2.0) | (1.9) | | (1.7) | (2.1) | (1.7) | | (1.7) | (1.8) | (1.6) | | (1.7) | 2.1 | (2.1) |

**Table 4-2 Field study data set showing number of runs, GM(GSD) for all runs grouped by location, direction of travel, season and time of day (GM - geometric mean, GSD - geometric standard deviation, cw - clockwise, acw - anti-clockwise)**

In order to assess the seasonal, diurnal and between route variability of the mean exposure values, an additional statistical analysis of the between group differences was carried out. This analysis was based on geometric mean values for all data collection runs, i.e. for each individual run all raw data was extracted and a geometric mean value was calculated. The geometric mean values from all runs in the associated sub-categories (e.g. 'summer' and 'winter') were compiled and an equality test of the mean was carried out, using the EViews function 'Test for Equality of Means Between Series' (QMS, 2002). The *p*-value returned by this function is based on the t-test. The findings from this analysis are outlined in the following sections.

### 4.2.1 Seasonal and diurnal variability

Results

According to the geometric mean values from Table 4-2, driver exposure to ultrafine particles is on average higher during morning rush hour (8 - 9 am) and during the winter months (April-September) compared to evening and summer runs, respectively. Average exposure values to $PM_{10}$ and $PM_{2.5}$ show the same diurnal variation, with measured concentrations being higher in the morning, but an inverse seasonal behaviour, i.e. summer concentrations of these particles appear to be higher than winter concentrations.

The analysis of geometric mean values for data from all runs showed that for ultrafine particles (UF) there is a statistically significant difference between summer and winter concentrations ($p < 0.0001$) while no such observation could be made for $PM_{10}$ and $PM_{2.5}$ mass concentrations (both $p > 0.07$). Similarly, a significant difference was found for the diurnal variation in geometric mean values of ultrafine particles ($p < 0.0001$), whereas differences in $PM_{10}$ and $PM_{2.5}$ values were less significant ($p = 0.002$ and $p = 0.011$, respectively).

Discussion - seasonal variability

The lack of statistical evidence regarding the difference between summer and winter mass concentration levels of $PM_{10}$ and $PM_{2.5}$ observed during the present study is in contrast to findings from Adams et al. (2001a), who reported significant seasonal variability of in-car exposure to $PM_{2.5}$ particles, with levels being generally higher during the summer months (35 $\mu g\ m^{-3}$) than the winter months (23.7 $\mu g\ m^{-3}$). It is thought that these seasonal differences are due to variations in meteorological

conditions. For example, Adams et al. (2001b) found that wind speed, temperature and relative humidity affect commuter exposure levels to $PM_{2.5}$ particles, with varying significance for winter and summer data sets. It is not clear why no significant difference was found between winter and summer mass concentration data in the current study. It is possible that this is a result of differences in the monitoring methods (i.e. optical devices are less accurate than the gravimetric methods) and/or vehicle setup (e.g. ventilation settings) used.

Results from the present study indicate that in-vehicle concentrations of ultrafine particles are generally higher during the winter period than the summer period. Although not directly comparable, this in accordance with findings from studies on traffic related particle number concentrations (not in-vehicle). For example, Kittelson et al. (2000) found that increased concentrations of nanoparticles (i.e. $< 0.05$ μm) were measured at lower ambient temperatures and Wehner and Wiedensohler (2003) report that higher concentrations of ultrafine particles were observed near roads during the winter months, which they linked to the effect of meteorological conditions on the dispersion of emissions. Moreover, ambient temperature is known to influence vehicle emissions due to the variation in engine efficiency. Emissions are particularly high for cold engines. As noted in Chapter 2, a report by AEAT (2001) states that measured emissions of ultrafine particles were higher for a cold engine (i.e. cold start emissions) than the corresponding hot start emissions. This is substantiated by findings from Alm et al. (1999) who report that especially high in-vehicle particle counts were recorded when driving behind buses and some petrol vehicles during the winter months. However, they also point out that average in-vehicle concentrations of particles in the range 0.3 - 1.0 μm are not significantly affected by season. This is contrary to results from the present study, which could be due to the difference in particle sizes investigated and the comparatively low number of sample runs ($N = 12$ for each season) used by Alm et al. (1999).

The findings regarding the seasonal variability indicate that the prevailing behaviour of in-vehicle concentration of ultrafine particles seems to follow the patterns for outside concentrations reported in other studies, and that their behaviour differs from that of larger particles and mass concentration values.

Discussion - diurnal variability

Results from the current study indicate that in-vehicle particle concentrations are generally higher during morning rush hour than during evening rush hour for both ultrafine particles and mass concentration values (although not significant for the latter at $p < 0.001$). This is in accordance with outside concentration measurements reported in several studies. For example, Molnár et al. (2002) observed more substantial increases in particle concentrations during morning rush hour than during evening rush hour as did Morawska et al. (1999) and Harrison et al. (1999a). These patterns were attributed to increased wind speeds, and thus dispersion, in the afternoon as well as a prolonged but lower traffic flow peak in the afternoon/evening compared to the more pronounced morning rush hour peak flow. However, particle formation processes also depend on meteorological conditions, resulting in elevated concentrations during the cooler morning hours, particularly for very small particles. For example, Kittelson et al. (2000) measured increased concentrations of nanoparticles at lower ambient temperatures, while Wehner and Wiedensohler (2003) linked peak number counts of particles in the range 10 - 15 nm to increased global radiation, which is thought to photochemically induce nucleation of gaseous components. As noted in the previous section, cold start emissions also contribute to elevated particle concentrations during periods with low ambient temperatures (AEAT, 2001).

Similar to the present study, findings from commuter exposure studies showed that this diurnal variability can also be observed for mass and number concentrations in commuter vehicles. Average $PM_{2.5}$ concentrations measured by Gómez-Perales et al. (2004) were generally higher during morning journeys than evening journeys. A significant diurnal difference was also found by Alm et al. (1999) for particle number concentrations, who reported increased mean concentrations of fine (0.3 - 1.0 µm) and coarse particles (1.0 - 10 µm) during morning journeys compared to afternoon journeys of up to 105 % and 60 %, respectively.

### 4.2.2 Between route variability

Results

It is also apparent from Table 4-2 that average concentrations on Route 2 (Abbey Lane) are generally higher than on the other two routes. This is true for ultrafine particles as well as mass concentration values. This observation is confirmed when

average values for all individual road links are considered. A table containing these values can be found in Table B-1 in the appendix. Figure 4-1 shows minimum and maximum geometric mean concentrations for the links on the three field study routes, separated into clockwise and anti-clockwise links. Ultrafine particle results are shown in blue, using the primary y-axis, while $PM_{10}$ and $PM_{2.5}$ values are plotted in yellow and red respectively, using the secondary y-axis.

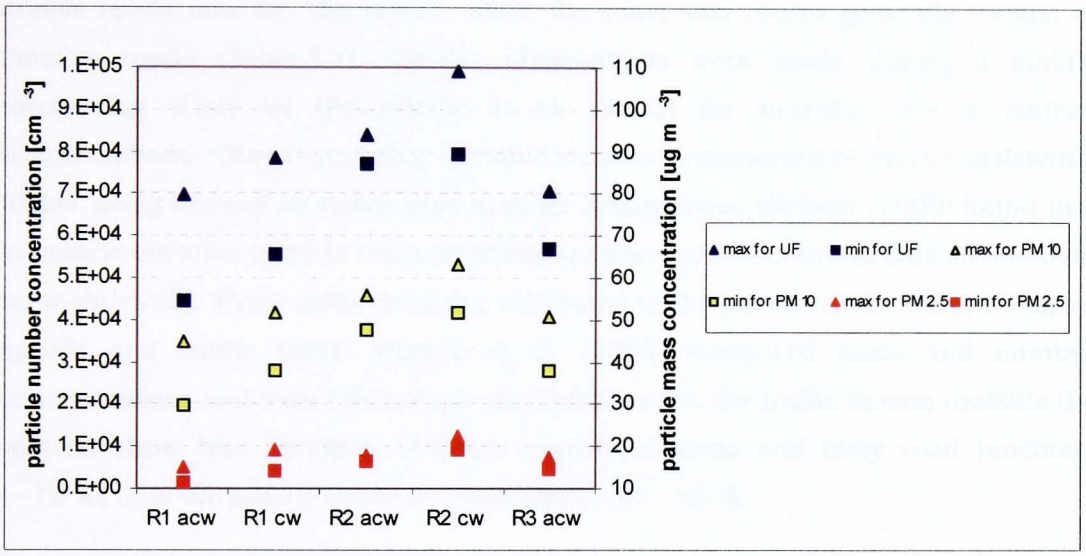


**Figure 4-1 Minimum and maximum geometric mean values for particle concentrations on the field study routes, separated into clockwise and anti-clockwise links**

The plots for all three particle ranges show similar trends. Although the ranges overlap in some cases, average particle concentrations are generally higher for the links on Route 2.

Statistical tests on geometric mean values for all runs confirmed that there is a statistically significant difference ($p < 0.0001$) in mean concentrations from Route 2 and each of the other two routes, for both ultrafine particles and $PM_{10}$. For $PM_{2.5}$ concentrations, however, the difference is not significant. No significant difference could be found between mean concentrations measured on Routes 1 and 3 ($p > 0.3$ for all particle fractions).

Discussion

A significant difference in average in-vehicle concentration values was found between one of the field study routes and the other two. Similar results were reported by

Adams et al. (2001b) who found that route was a significant factor, explaining approximately 20 % of the variation in commuter exposure levels (including cyclists, bus and car commuters). The observed effect of the parameter 'route' on differences in exposure clearly depends on the routes investigated. In the present study, Route 2 was associated with higher exposure levels, while no significant difference was found between results for Route 1 and 3. This is likely to be due to Route 2 consisting mainly of dual carriage type roads which are part of the inner ring road and other main access roads into the city centre, while the other two routes generally consist of smaller roads (Table 3-1). Similar observations were made during a mobile monitoring study by Bukowiecki et al. (2003) for in-traffic (not in-vehicle) concentrations. They report that ultrafine number concentrations were consistently higher along freeway stretches than in other urban areas. Dickens (2000) found that in-vehicle ultrafine particle concentrations are lower on small urban side streets than on main roads. These differences are attributed in the papers to the effect of higher speeds and traffic flows. Weijers et al. (2004) compared mass and number concentrations and found that mass concentrations in the traffic stream (outside the vehicle) show less variation between residential areas and busy road junctions (~12 %) than ultrafine particle concentrations (40 - 83 %).

More detailed investigations are required in order to identify what characteristics make a route a 'high exposure' route. For example, a study by Alm et al. (1999) found that, on a fixed journey, low average journey speeds and the number of stops (e.g. at traffic lights or in traffic jams) could be associated with higher particle number concentrations (0.3 - 10 μm). Based on these observations, increased driver exposure would be expected to occur on routes which are more likely to be congested.

### 4.2.3 Summary

A preliminary analysis of the data has shown that particle concentrations inside the vehicle exhibit similar patterns to those found in other studies for outside concentrations. Ultrafine particles and mass concentration values behave similarly in terms of diurnal and route specific variation, with higher values recorded during the mornings and on routes with increased traffic flow and higher speeds.

However, there are also significant differences in the patterns exhibited by the three particle size fractions with regard to seasonal and day-to-day variations. Number

concentration values tend to be higher during the winter months while there are indications that mass concentration values are slightly increased during the summer months.

In order to investigate in more detail the degree to which external parameters are linked to in-vehicle concentration values, a multiple regression analysis was carried out to identify the main determinants of driver exposure. This is reported in Chapter 5. In the first instance however, it is instructive to (a) further explore the field study data in order to uncover any potential relationships between the measured parameters (e.g. speed, particles from different size ranges) and (b) to derive a general understanding of the short-term and long-term variability of driver exposure. The next section focusses on time-lag investigations and is followed by a description of various data visualisation techniques suitable for viewing large quantities of driver exposure data.

## 4.3 Temporal dynamics of driver exposure

Following the investigation of average driver exposure values in the previous section, the methods used in this section aimed to make use of the high-resolution character of the data to investigate the short-term variability of driver exposure to particulate matter.

In the first instance, traditional time series plots were used to investigate the characteristics of time series data for the different particle size fractions and to examine the effect of data smoothing.

However, with a large number of individual data sets, e.g. from separate data collection runs, it proved to be impracticable to make visual comparisons of time series traces, particularly since the aim was to identify similarities and differences between related data sets. It was therefore considered useful to investigate ways of visualising time series data that would facilitate the inspection of multiple plots and would thus allow for visual inspection of underlying trends and patterns. A suitable visualisation technique was developed and used to examine time series plots for single and multiple runs.

The techniques used and the findings derived from the different plots are described in the following sections.

### 4.3.1 Characteristics of time series data

The visual inspection of time series plots for the field data set revealed that ultrafine particles show distinct periods of high and low concentrations while mass concentration traces ($PM_{10}$ and $PM_{2.5}$) have a higher short-term variability.

Figure 4-2 shows an example for a typical time series of in-vehicle concentration data for the three size fractions. The black lines in the figures, representing the raw measured data, illustrate the different variability of number (UF) and mass concentrations ($PM_{2.5}$ and $PM_{10}$) measurements.

The extreme short-term variability exhibited by the mass concentration traces is thought to be partly due to these values being based on the detection of relatively few particles, i.e. a change in the number of particles detected has a strong effect on the measured value, leading to high second by second variability. It is possible that this effect is initially small but is further amplified by turbulences generated when outside air enters the vehicle (e.g. through ventilation or windows) and mixes with the in-vehicle air. Movements of the driver may also affect turbulence. Such microturbulence may be a contributing factor for the high variability of in-vehicle concentrations, which is observed here due to the high temporal resolution of the measurements. This is supported by the observation that concentrations of ultrafine particles also show second to second variability, though on a smaller scale.

By removing all periodicities with relatively high frequencies from the time series, a smoothing effect can be achieved which aids the detection of 'real' peaks, i.e. elevated concentrations that persist over extended periods of time. A Fourier transformation with low pass filter was found to be a useful tool for this purpose. Figure 4-2 shows the effect of data smoothing on number and mass concentration traces (i.e. red line). The Fourier cutoff value used here is 10 s (i.e. all periodicities > 1/10 Hz are removed), which is thought to be sufficiently short to remove extreme short-term variability but still retain adequate detail in order to detect short transient events such as a car in front or a short acceleration period. Visual inspections of multiple time series from different data collection sessions confirmed that cutoff settings between 10 and 20 s best aid the detection of peaks while removing uninformative variation in the data. Although this smoothing technique was found to be useful for the initial investigation of time series traces, the original 1 Hz time series data sets were generally used for the analysis.

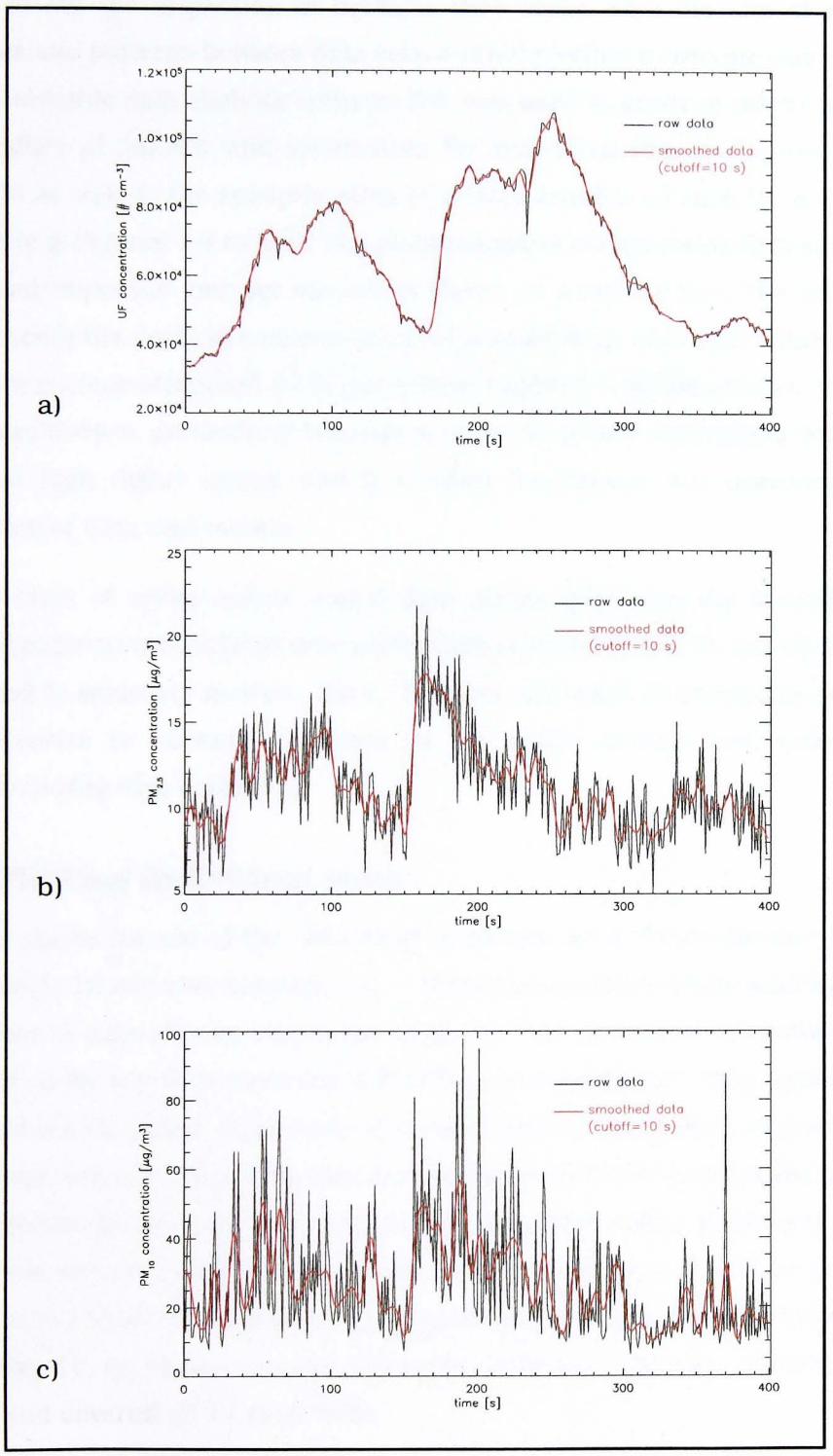


**Figure 4-2 Example time series of concentrations of ultrafine particles, $PM_{2.5}$ and $PM_{10}$, showing raw data (black) and Fourier smoothed data lines (red)**

### 4.3.2 Novel plotting technique

In order to aid the inspection of multiple time series with the aim of identifying similarities and patterns between data sets, a novel plotting technique was developed. The progammable data analysis software IDL was used to produce colour coded time series profiles of particle and speed data for individual data collection runs (e.g. Figure 4-3) as well as the synoptic plots of a large number of runs for each particle size range (e.g. Figures 4-4 to 4-6). The plots consist of colour coded data strips where each measurement (i.e. one per second) is shown as a vertical line. The colour of the line represents the particle concentration on a scale from blue (low values) through red (medium concentrations) to bright yellow (highest concentrations). The colour scheme was chosen particularly because it helps to clearly distinguish between low (dark) and high (light) values, and it is used throughout this document for the visualisation of data and results.

This approach of using colour coded data strips facilitates the identification of significant patterns across large time series data sets which may be lost when the data are reduced to summary metrics. Here, the plots were used to investigate patterns in driver exposure in relation to events in the traffic stream and meteorological conditions during data collection.

### 4.3.3 Plotting individual runs

Figure 4-3 shows the use of the 'data strip' technique for a data collection journey of approximately 10 minutes duration, i.e. ~ 600 concentration values were logged. The plot consists of time-aligned horizontal strips for four measured quantities: number concentration for ultrafine particles (UF); $PM_{2.5}$ concentration; $PM_{10}$ concentration; and instantaneous speed. Each strip is formed from a continuous sequence of thin (i.e. 1 second 'wide') vertical lines that have been colour-coded to show the magnitude of the respective parameters. The vertical gray lines that extend above and below the strip indicate when the vehicle passed from one link to the next. The time series strips shown here are based on data from one completion of Route 1 during morning rush hour on May 16, on Route 1 in anti-clockwise direction. The data collection started on link 7 and covered all 11 road links.

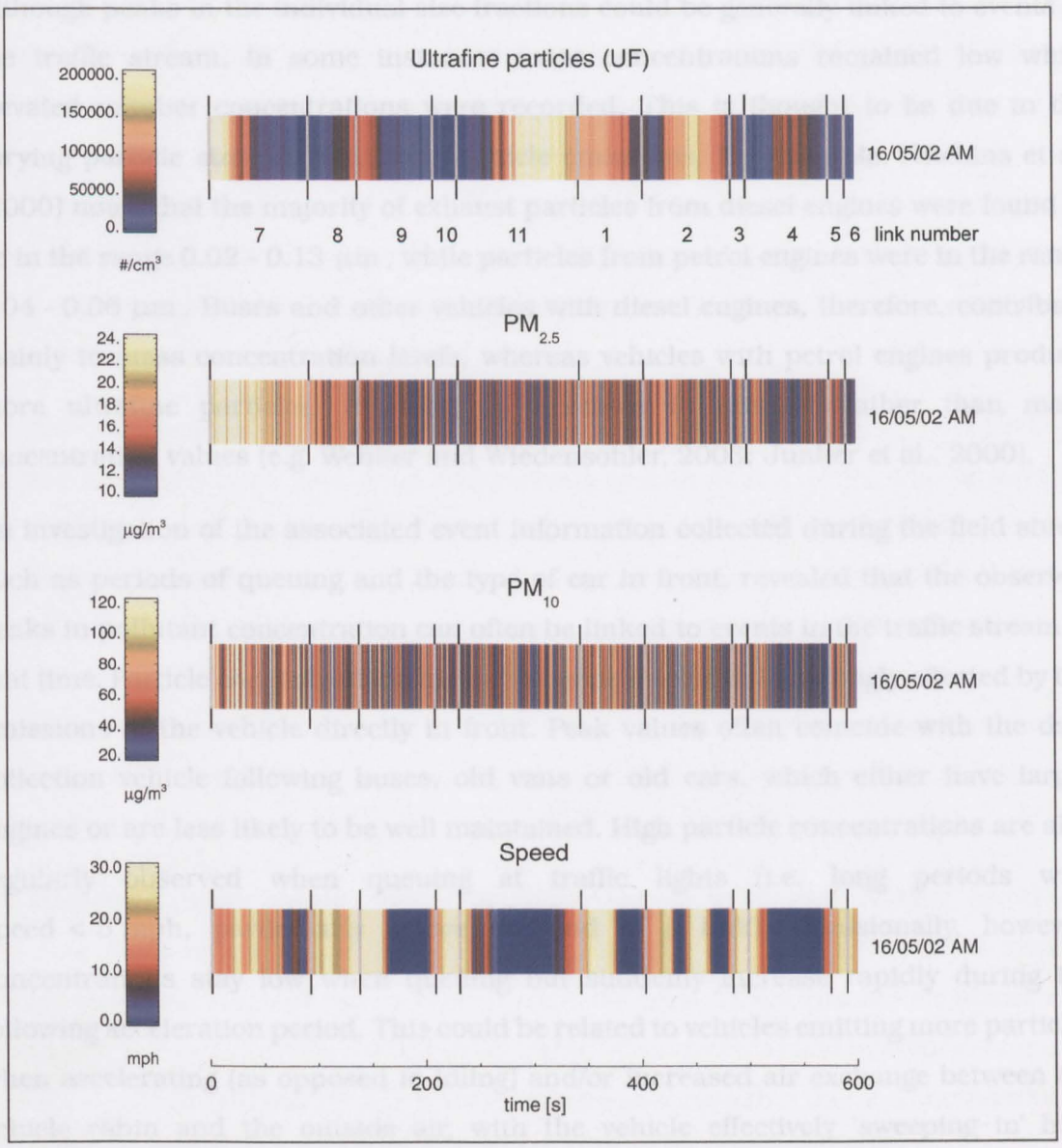


**Figure 4-3 Example plot showing particle and speed data strips for data collected during one completion of Route 1, AM, anti-clockwise, logging interval: 1 second**

The speed strip vividly conveys the kinematics of the vehicle: short periods of acceleration (changing colour from blue through to yellow); steady speeds around 30 mph (yellow); and long stationary periods (blue). The most readily apparent association between parameters is the sustained high ultrafine particle concentration occurring during the long period of queuing (i.e. stationary) on link 11.

As observed in the time series plots in Section 4.3.1, ultrafine particle count data clearly show distinct periods of high and low concentrations whereas $PM_{10}$ and $PM_{2.5}$ data show more short-term variability. However, the time series plots for ultrafine data

clearly show different patterns than the ones for mass concentration measurements. Although peaks in the individual size fractions could be generally linked to events in the traffic stream, in some instances mass concentrations remained low while elevated number concentrations were recorded. This is thought to be due to the varying particle size distribution of vehicle emissions. For example, Hitchins et al. (2000) notes that the majority of exhaust particles from diesel engines were found to be in the range 0.02 - 0.13 μm, while particles from petrol engines were in the range 0.04 - 0.06 μm. Buses and other vehicles with diesel engines, therefore, contribute mainly to mass concentration levels, whereas vehicles with petrol engines produce more ultrafine particles, resulting in increases in number rather than mass concentration values (e.g. Wehner and Wiedensohler, 2003; Junker et al., 2000).

An investigation of the associated event information collected during the field study, such as periods of queuing and the type of car in front, revealed that the observed peaks in pollutant concentration can often be linked to events in the traffic stream at that time. Particle concentrations inside the vehicle seem to be strongly affected by the emissions of the vehicle directly in front. Peak values often coincide with the data collection vehicle following buses, old vans or old cars, which either have larger engines or are less likely to be well maintained. High particle concentrations are also regularly observed when queuing at traffic lights (i.e. long periods with speed < 5 mph, particularly before the end of a link). Occasionally, however, concentrations stay low when queuing but suddenly increase rapidly during the following acceleration period. This could be related to vehicles emitting more particles when accelerating (as opposed to idling) and/or increased air exchange between the vehicle cabin and the outside air, with the vehicle effectively 'sweeping in' high pollutant concentrations that have accumulated in the ambient air during the queuing period.

The time series data in Figure 4-3 illustrate these instances, some of which are described in detail as follows. According to the event log, the high concentration episode at the beginning of the run ends shortly after a relatively old car that was being followed turned off. Elevated concentrations values can be observed for both mass and number concentrations. The number concentrations seem to level out quickly while mass concentration values decrease only slowly. The elevated particle number concentrations recorded at the end of link 8 and beginning of link 9 coincide with an

episode of high speed directly after an old taxi has been followed for only 16 seconds. Particle mass concentration values seem unaffected at this point. During a long queuing period on link 11, a gradual build up of ultrafine particles followed by a high concentration episode (link 1) was recorded. A similar, though less severe, build up occurred again later on link 4. Those two events did not affect particle mass concentrations ($PM_{2.5}$ and $PM_{10}$). Another particle number concentration peak was recorded on link 2. This cannot be attributed to any particular vehicle being in front, but it does coincide with a short period of high speed. No clear effect on particle mass concentrations during this time could be observed.

Similar observations regarding the effect of the vehicle directly in front on in-vehicle particle concentrations have been documented by Alm et al. (1999), who found that peak number concentrations (particle size 0.3 - 10 μm) were recorded inside the test vehicle when driving behind city buses. Cold gasoline-powered cars were also linked to particularly elevated in-vehicle concentrations. Comparative results were published by Dickens (2000) who identified buses as having the strongest effect on ultrafine particle concentrations inside the vehicles following them. However, some individual cars (and taxis) were found to be worse, even though in general cars were linked to lower in-vehicle concentrations than buses and other diesel powered vehicles.

As Dickens (2000) points out, it is important to note that linking in-vehicle concentrations to the vehicle directly in front can only give indicative results since it is not clear how much of the measured particles comes from the vehicle ahead and how much is emitted by the surrounding traffic. Experimental and analytical research into the dispersion of particles in vehicle wakes has shown that this is a complex issue which is not yet fully understood (e.g. Hider et al., 1997; Baker, 2001).

Time series and event data from the present study indicate that queuing periods also affect in-vehicle concentrations. Although similar findings have not been published for in-vehicle studies, results from Weijers et al. (2004) indicate that these elevated concentrations are linked to variations in outside concentration. They report that emissions directly outside the vehicle in congested traffic are subject to extreme fluctuations, induced by accelerating and decelerating of high numbers of vehicles. They measured increases in mass concentrations of approximately 30 % and in ultrafine particle concentrations of 400 % during periods of congestion. Although these values are based on outside concentration measurements, it is likely that these

high concentrations lead to increased in-vehicle concentrations, especially for ultrafine particles, as observed in the present study.

Despite the uncertainty regarding the processes and time-scales involved of the ingress of particle emission from the surrounding traffic into the vehicle, the association of event data with in-vehicle concentrations helps to gain an understanding of the complexity of the factors affecting driver exposure.

### 4.3.4 Synoptic plots for multiple runs

It is instructive to be able to view colour coded data plots for multiple time series data sets simultaneously. These charts are referred to as synoptic plots. Figures 4-4 to 4-6 show examples of such synoptic time series plots for data from 21circuits completed on Route 1 during morning rush hour from April to August 2002. The data shown in Figure 4-3 is included also. These plots vividly convey the degree and character of the variability in exposure for one route on different days.

Figure 4-4 shows instantaneous speed data while Figures 4-5 and 4-6 show data strips of particle concentrations for the ultrafine and $PM_{2.5}$ ranges, respectively. Additional information about the data collection date and the value for the total exposure experienced are included for each run. Link divider lines are included in all plots in light green but, due to scale, can be difficult to distinguish from the data.

Speed data

As is to be expected, the time taken to complete the same circuit varied for each journey. Thus, the length of the (blue through yellow) shaded part of the strips is proportional to the duration of the journey (the remainder where there is no data is shaded gray). The shortest time to complete this circuit was approximately half that for the longest time. The speed data shows clear patterns, indicating that traffic lights and speed limits lead to vehicles following similar drive cycles. The ends of the speed data bars at the bottom half of the plot, for instance, have similar patterns of queuing up (at lights), accelerating and driving at higher speed on the last two short links. A reason for the upper half of the plot not showing the same patterns is believed to be the school holidays which affect traffic density, particularly in the proximity of schools such as the ones near links 5, 6 and 7. Also, the journeys are generally shorter during school holidays.

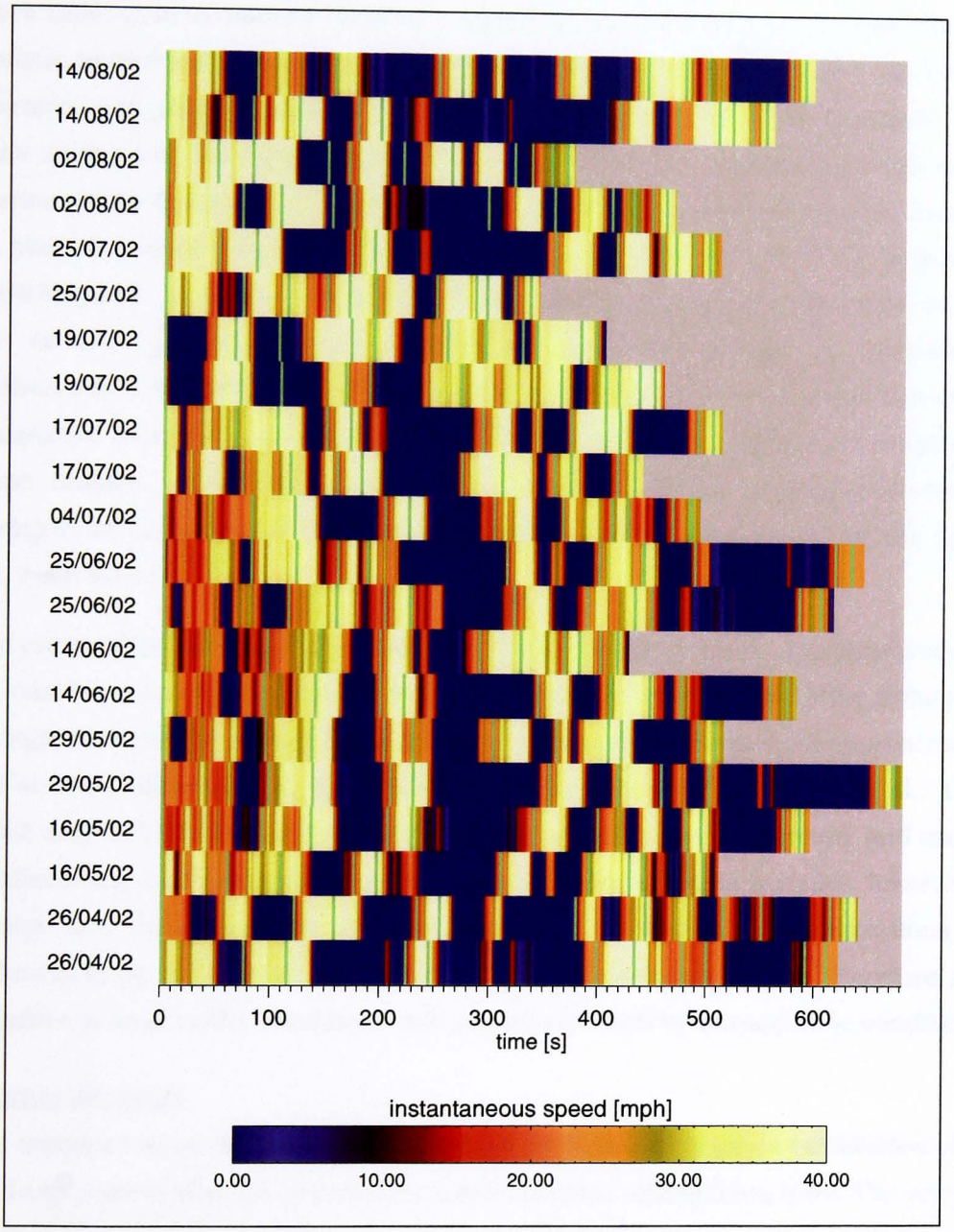


**Figure 4-4 Instantaneous speed - synoptic plot for time series data from multiple journeys**

*Route 1, 21 morning rush hour runs, anti-clockwise direction, logging interval: 1 second*

Particle concentrations

Similar to the plots in Figure 4-3, ultrafine particle concentrations show distinct peaks, often extended over one or more whole links, whereas the $PM_{2.5}$ plots show more short-term variability. However, examining multiple journeys, it is observed that particle concentrations not only vary within one time series (journey) but also between journeys. The particle mass concentration plot (Figure 4-6) shows a pattern of two runs carried out during the same data collection session as having generally similar concentration levels. For example, concentrations stayed high throughout both runs on May 16 and also on June 25. Other days such as April 26 and May 29 generally show very low values apart from some obvious peaks. It is thought that these patterns are caused by variations in atmospheric conditions which, as discussed in Section 2.2.1, can affect both background concentrations and the dispersion of local emissions. Whereas local emissions are likely to vary between different runs from the same session, depending on the vehicles in the vicinity during measurement, background conditions are expected to remain virtually constant during the session (i.e. rush hour).

The observation that these patterns are more apparent for mass concentrations than for number concentration values indicates that they may be due to the influence of background concentrations. Large particles which contribute to mass concentrations, are strongly influenced by background concentrations (e.g. Harrison et al., 1999), which may be high during inversion periods, due to reduced dispersion, and can also be affected by long range transport (APEG, 1999). For ultrafine particles, however, the background concentrations are low compared to in-traffic concentrations (e.g. Bukowiecki et al., 2003). Measured number concentrations are therefore more sensitive to local traffic conditions and are less affected by atmospheric conditions.

Journey exposure

The exposure values at the right hand side of each data strip give an indication of how total exposure is affected by pollutant concentrations and journey time. The exposure experienced during a short journey with relatively high particle concentrations may by far exceed what is experienced during a long journey at low concentrations. Average exposure values range from 8 - 22 $\mu g\, m^{-3}$ for $PM_{2.5}$ and 20 - 130 $x10^3\ cm^{-3}$ for particle counts.

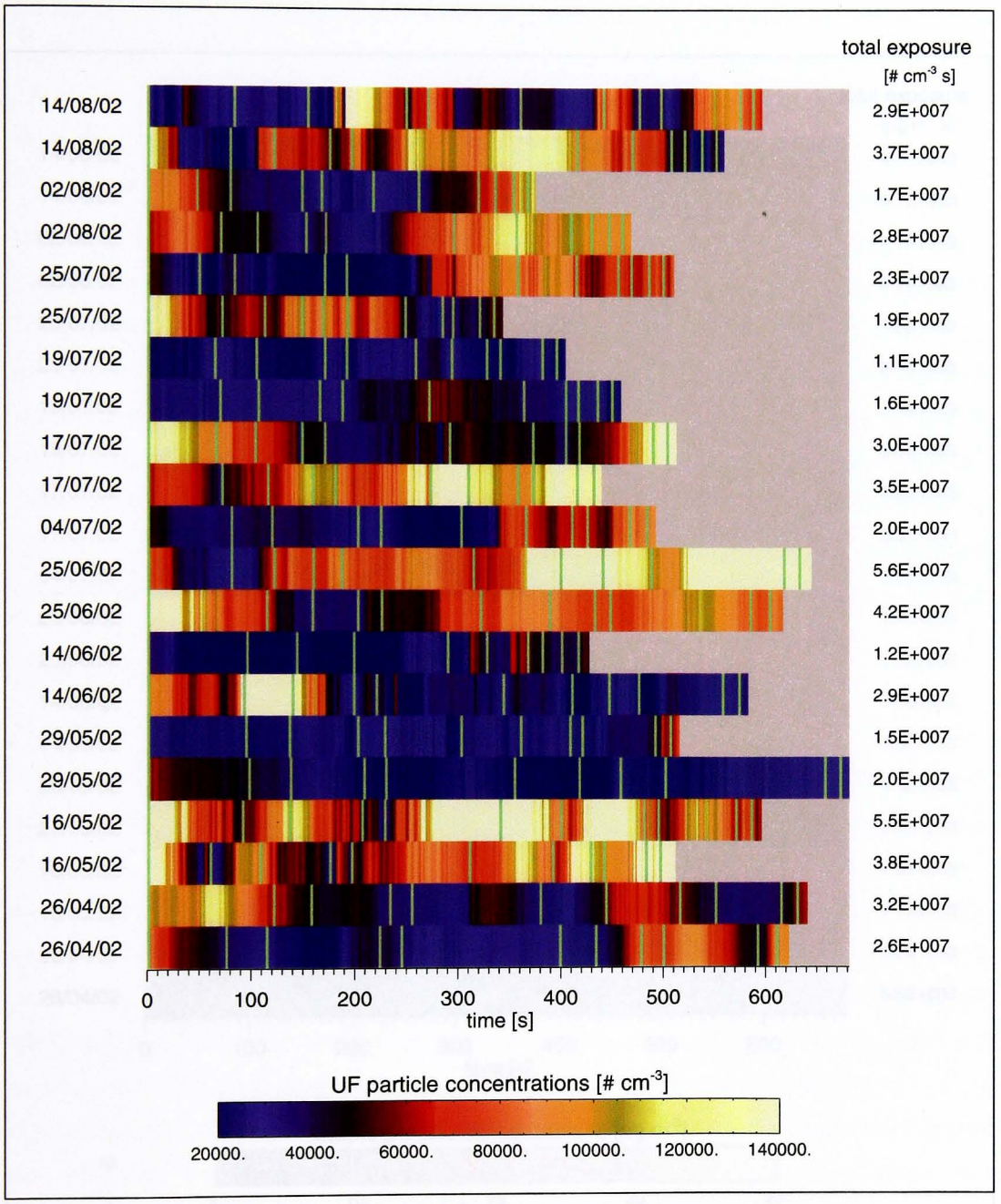


**Figure 4-5 Ultrafine particle number concentration - synoptic plot for time series data from multiple journeys**

*Route 1, 21 morning rush hour runs, anti-clockwise direction, logging interval: 1 second*

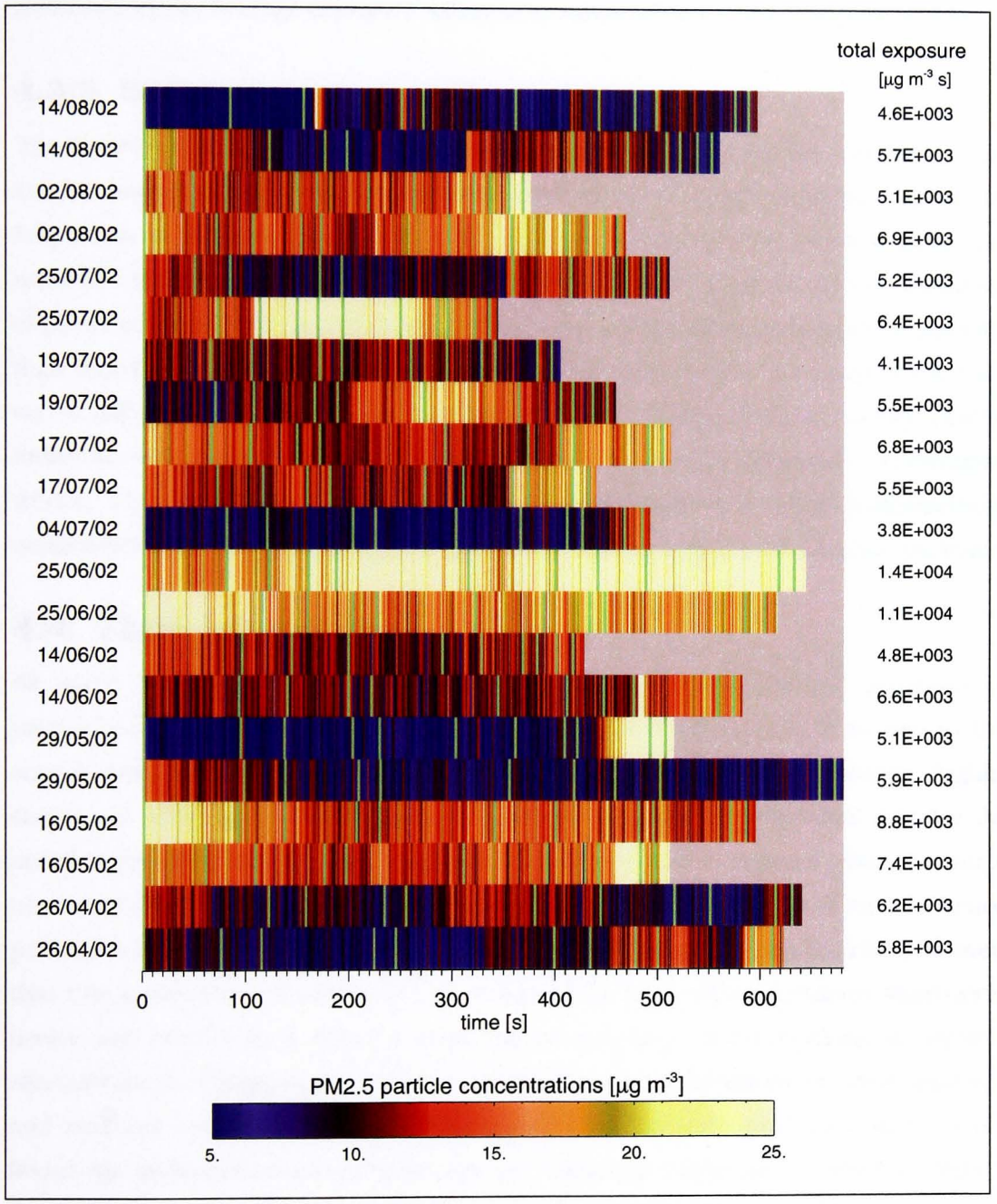


**Figure 4-6 $PM_{2.5}$ mass concentration - synoptic plot for time series data from multiple journeys**

*Route 1, 21 morning rush hour runs, anti-clockwise direction, logging interval: 1 second*

Comparable plots were examined for clockwise and evening runs on Route 1 as well as runs for Routes 2 and 3. They all show the described features, i.e. peaks that can be linked to events in the traffic stream and runs from the same day exhibiting similar pollution levels. Average exposure values are comparable to the ones given above.

### 4.3.5 Summary

The described plots of field study data indicate that in-vehicle particle concentrations can be associated with events in the traffic stream and to meteorological conditions. Inspection of the plots did not reveal any obvious correspondence between mass and number concentration profiles. The plots indicate that the mass concentration values of particles show different short-term variability and dependence on traffic emissions than number concentrations. Investigations of the short-term variability of the time series data showed that elevated concentrations of ultrafine particles, which can be linked to events in the traffic stream, are often not matched by the mass concentration traces. These findings further support the hypothesis that mass concentration measurements of $PM_{10}$ and $PM_{2.5}$ cannot be used as surrogates for ultrafine particles.

## 4.4 Time-lag analysis

As noted in Section 2.2.1, it is generally accepted that the rate of emissions of particulate matter from vehicles depends on the speed of the vehicle as well as the acceleration. A faster moving stream of traffic may therefore generate higher emissions, which in turn will affect concentrations inside the individual vehicles. As noted in Section 2.3.2, it is not clear exactly how ventilation regimes affect in-vehicle concentrations. Moreover, research has shown that the penetration of air and hence pollutants into the vehicle depends on speed also. For example, Clarke (1998) showed that this penetration process acts as a time-delay filter which removes short-term peaks and results in a delayed response of in-vehicle concentrations to outside concentrations. Chang et al. (2000), who investigated correlations between personal and ambient concentrations of $PM_{2.5}$, report that the strongest correlations were found for microenvironments with high air exchange rates, i.e. in vehicles. Higher particle concentrations inside the vehicle may therefore be expected to occur at higher speeds due to an increased ingress of potentially higher outside concentrations into the vehicle.

In order to investigate whether this effect can be identified in the field study data, a cross-correlation analysis was carried out. The correlation coefficient for the cross-correlation between instantaneous speed and particle concentrations was calculated for various time-lags, using the C_CORRELATE function in IDL (see Eq B-1). To reduce the chance of the high resolution of the raw data (i.e. logged at one second intervals) masking any underlying patterns, additional time series were calculated for a number of longer averaging intervals (i.e. 2 s, 5 s, 10 s, 15 s, 20 s, 25 s, 30 s, 35 s and 40 s). This was achieved by averaging raw per second data over the 'new' interval and compiling the results in a new time series. This method was intended to simulate measuring at logging intervals > 1 second, i.e. one data value is logged for each time interval based on the average of all individual concentration values measured during the interval. Time-lag correlations were then carried out for the original data as well as the modified time series. Since the field study data consists not of one continuous time series but rather of a number of individual runs, lasting from 4 to 30 minutes, this method was applied separately to each section of continuous data (e.g. each complete run carried out on a route).

Cross-correlation results were found to vary strongly between different runs. It was therefore decided to aggregate the cross-correlation results for all runs in order to investigate whether any underlying tendency of higher correlation values at a particular averaging interval could be identified. This was achieved by calculating the cross-correlation for each run and averaging interval and compiling the results for each of the three field study routes. Using these compiled results, a number of time-lag matrices were generated. The matrices allow an inspection of aggregated cross-correlation values for instantaneous speed and particle concentrations for all runs carried out on each route. Each matrix shows the distribution of the cross-correlation values for data from a particular route and averaging interval, depending on the time-lag. Examples for cross-correlation matrices for raw data and data averaged over 20 seconds are shown in Figure 4-7.

The matrices have been generated by using a histogram like approach, i.e. counting the number of times a particular cross-correlation value occurred for a particular time-lag and assigning this result to a bin. The colour of the bin in the plot represents the number of values recorded for this combination of time-lag and cross-correlation. Empty bins are coloured grey. The y-axis of each plot shows the cross-correlation

value and the x-axis shows the time lag used in the calculation. In plots a to c the unit of the time lag axis is in seconds. Since plots d to f, however, are based on time series data where each value represents a 20 second average, the time lag is more like a factor, i.e. a lag of 10 representing an actual difference of 200 seconds. The fact that plots d to f have less values for longer time-lags, compared to plots a to c, is due to the time series for 20 second averages being more likely to be too short to allow cross-correlation calculations for these lag values.

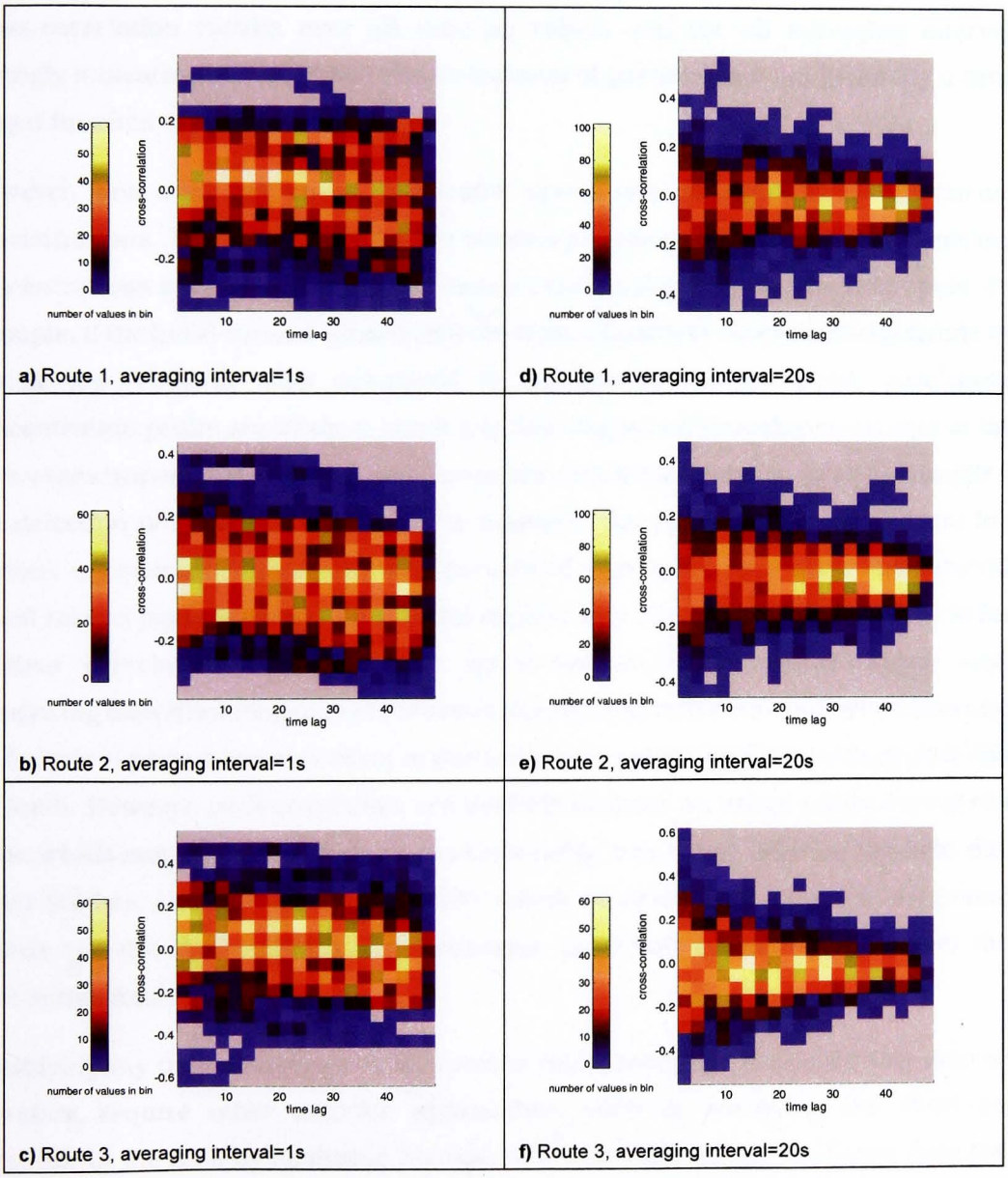


**Figure 4-7 Cross-correlation matrices for speed vs. ultrafine particle concentrations for Routes 1 to 3, using time series data logged at intervals of 1 second (a-c) and averaged over 20 seconds (d-f)**

The plots show that no obvious lagged relationship between instantaneous speed and in-vehicle concentrations of ultrafine particles could be detected. The cross-correlation values are typically rather low, i.e. between -0.2 and 0.2. Occasionally, higher values are observed but they are very infrequent. Although on the occasional run there may appear to be a dependence of in-vehicle concentrations on instantaneous vehicle speed, this cannot be confirmed generally for other runs. In fact, the plots show that occasionally higher cross-correlation values are observed for all time-lags and in both directions, positive and negative. This rather even spread of cross-correlation results over all time-lag values and for all averaging intervals strongly indicates that in-vehicle concentrations of particles are not generally a time-lagged function of speed.

However, this does not mean that traffic speed has no effect at all on particle concentrations. The time-scales for the various processes affecting in-vehicle particle concentrations may be so diverse that causal relationships may not be detectable. For example, if the build-up and subsequent decrease of particle concentrations inside the vehicle are relatively slow compared to changes in vehicle speed, overlapping concentration peaks are likely to mask any time-lagged relationship to an extend that it becomes impossible to detect. In some cases, conditions may be more favourable to the detection of such relationships. For example, the comparison of data from long periods of constantly low speed with periods of high speed may allow the detection speed related particle peaks. This would explain why Dickens (2000) was able to find a 'clear correlation' between traffic speed and in-vehicle concentrations when comparing data from long periods of motorway driving with rural and urban journeys. He further reported that this effect is particularly pronounced for speeds greater than 50 km/h. However, such conditions are unlikely to occur on urban roads during rush hour, which may explain why no such relationship was found here for the field data. Other studies, such as Alm et al. (1999), which reported links between low journey speeds and increased particle concentrations, used only average values rather than time series data.

Identifying any potential speed-concentration relationships in the field study data will therefore require other analysis approaches, such as removing the short-term variation in the data by analysing average values or inspecting time series data from individual runs in detail, in order to link individual peaks to events in the traffic

stream. In order to facilitate the latter, various visualisation techniques were explored, as described in Section 4.3.

## 4.5 Correlation analysis

The time series plots in Sections 4.3 have shown that the short-term variability of particle concentrations varies greatly for the different size ranges. They indicate that particles of the larger size ranges may not be suitable as surrogates for ultrafine particles. However, if the different size ranges are indeed correlated in some manner, the relations may not be obvious from raw data plots. In order to investigate whether any underlying correlations could be revealed by applying statistical methods to a large number of time series from the field study, a detailed correlation analysis was carried out.

All continuous data sections collected on the individual links were extracted from the data set. For each section (N = 308), the correlation coefficient $r^2$ was calculated for the three comparison cases: ultrafine vs. $PM_{2.5}$; ultrafine vs. $PM_{10}$; and $PM_{2.5}$ vs. $PM_{10}$. The average of these correlation coefficients for all three cases are shown in Table 4-3, together with the standard deviation.

| | UF vs. $PM_{2.5}$ | UF vs. $PM_{10}$ | $PM_{2.5}$ vs. $PM_{10}$ |
|---|---|---|---|
| mean $r^2$ | 0.155 | 0.069 | 0.229 |
| standard deviation | 0.159 | 0.090 | 0.182 |

**Table 4-3 Mean $r^2$ and standard deviation for correlation between particle size ranges (all $r^2$ significant at $p < 0.001$)**

These results show that, on average, the correlations between the time series data for ultrafine particles and $PM_{2.5}$ are rather weak and that the correlations between ultrafine particles and $PM_{10}$ data are even weaker. In both cases, the standard deviation is larger than the average of the correlation coefficient, which suggests that the correlations between time series data from these size ranges vary severely for different data sets. Stronger correlations, with less variability, were found for $PM_{2.5}$ vs. $PM_{10}$, which is to be expected since data for these size ranges are similar, i.e. they are both based on mass concentration measurements which include mainly particles with a diameter > 1 $\mu m$, as opposed to number concentrations which are dominated by particles < 1 $\mu m$.

However, it was considered likely that the failure to detect stronger correlations between ultrafine particles and the other size ranges may be due to bias, introduced by the difference between the short-term variability of mass and number concentration measurements. The short logging time of 1 Hz which was used for the measurements may make this bias even more pronounced. The individual time series data sets were therefore modified based on the following considerations.

1) The short logging interval of one second, on which the time series data is based, together with the extreme short-term variability in the PM data, makes it difficult to match peaks in the different time series, which may mask underlying correlations. A smoothing method was therefore applied to the data, using fast Fourier transformation where all periodicities below a certain cut-off value (e.g. 5 seconds) are removed, leading to a smoother data line. For further details on the methods used see Section 4.3.1.

2) Moreover, the short logging interval of one second, which was used for the measurements, may not be an ideal time-scale for this analysis. The use of a longer logging interval may have reduced the strong variability and given a clearer picture of the concentration behaviour and underlying correlations. In order to simulate longer logging intervals, the time series data was averaged over various time intervals from 5 to 20 seconds. This method was intended to emulate the operation of the monitoring equipment at logging intervals > 1 s, i.e. where a concentration value is logged for each measurement interval based on the average of all values measured during that interval.

Figure 4-8 illustrates the effects of longer averaging times and smoothing settings (i.e. Fourier cutoff) on the correlation between time series data for $PM_{10}$ and $PM_{2.5}$, using data from one typical journey. In this case, the correlation coefficient for the unmodified time series (Figure 4-8a) is rather low ($r^2 = 0.16$). By using smoothing settings (Figure 4-8b) or a longer averaging time (Figure 4-8c), the extreme short-term variability in the data can be reduced, leading to the detection of stronger underlying correlations ($r^2 > 0.36$ and $r^2 > 0.48$ respectively). By applying both settings, the effect can be increased further, resulting in an even higher correlation. It is important to note however, that high settings for smoothing or averaging interval may lead to a flattening of the time series traces, which may result in the apparent detection of strong correlations, artificially introduced by the smoothing process. Visual

inspections of multiple time series plots indicated that averaging times exceeding 20 seconds and smoothing cutoffs above 10 are unsuitable for this analysis due to the strong skewing of the data.

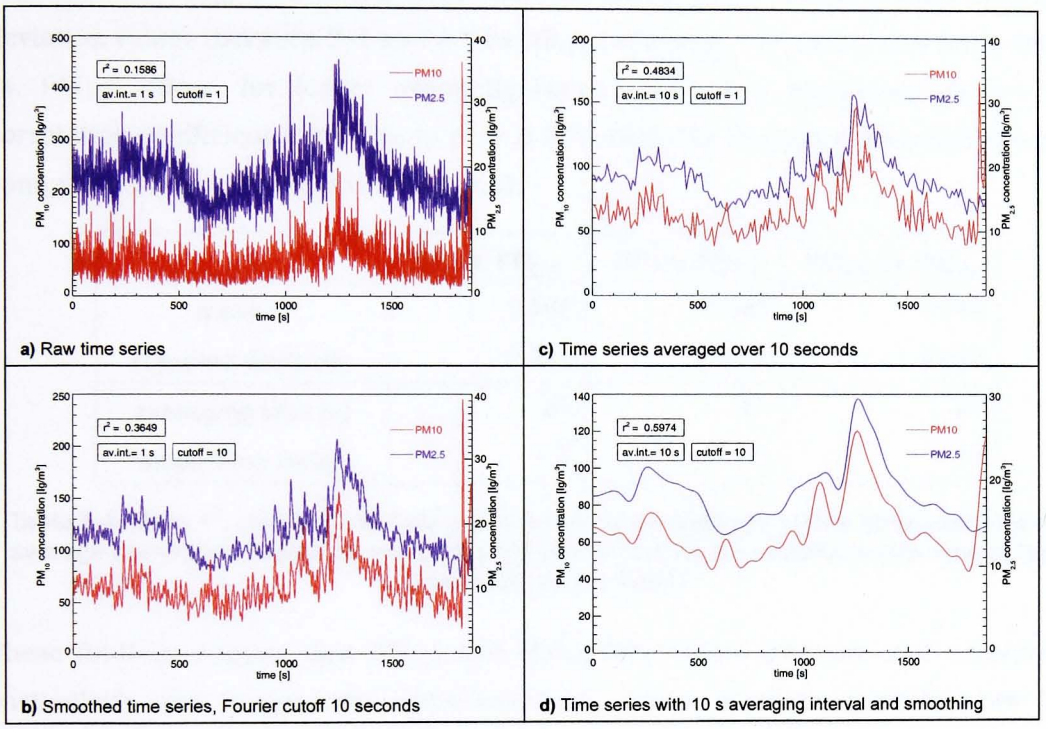


**Figure 4-8 Time series plots showing the effect of changes in averaging interval and smoothing settings**

In order to investigate which settings below these limits would be most suitable for the detection of underlying correlations, correlation coefficients were calculated for a number of different settings for averaging time and smoothing value. Colour coded matrices were generated which allowed the visual inspection of the distribution of correlation coefficients depending on these settings (see Figure B-1). For further details regarding the methods and data used to generate these matrices refer to Appendix B.3.

Using longer averaging times and applying smoothing techniques generally resulted in stronger average correlations in all three comparison cases. The averaging and smoothing settings which lead to the highest correlation coefficients are shown in Table 4-4.

For all three cases, higher mean correlation coefficients were found for longer averaging times. Smoothing settings also positively affected the detection of correlations. Although modifying the original time series data in this way resulted in stronger correlations being observed, the higher average correlation coefficients for ultrafine vs. $PM_{10}$ and $PM_{2.5}$ were also associated with similarly high standard deviation values (between 2.2 and 2.7 for all parameters). This is in contrast to $PM_{2.5}$ vs. $PM_{10}$, where, for longer averaging times and higher smoothing factors the correlation coefficient increases to $r^2 = 0.747$ while the standard deviation remains comparatively low at approximately 0.22.

| | UF vs. $PM_{2.5}$ | UF vs. $PM_{10}$ | $PM_{2.5}$ vs. $PM_{10}$ |
|---|---|---|---|
| **mean $r^2$** | 0.264 | 0.245 | 0.747 |
| **standard deviation** | 0.241 | 0.227 | 0.223 |
| **averaging time [s]** | 20 | 20 | 20 |
| **smoothing factor** | 5 | 10 | 10 |

**Table 4-4 Mean $r^2$, standard deviation and modification settings for the strongest detected average correlation based on modified time series data for three particle size ranges (all $r^2$ significant at $p < 0.001$)**

These findings suggest that $PM_{2.5}$ and $PM_{10}$ time series data are well correlated, particularly for longer-term measurements, while ultrafine particles are not consistently correlated with the larger size fractions. This is illustrated in Figure 4-9 for two sets of UF and $PM_{2.5}$ time series data. The correlations between the unmodified time series is reasonably good for the example shown in Figure 4-9a, but very weak for the example in Figure 4-9b. The application of smoothing settings increases the correlation for the first example from 0.33 to 0.45 (Figure 4-9c) while there is not much change for the second example (Figure 4-9d). The different results for the two examples illustrate how the correlation between ultrafine number concentrations and mass concentrations can vary, resulting from the different sensitivity of the size fractions to traffic events. As explained in Section 4.3, patterns in ultrafine concentrations are often not matched by the $PM_{2.5}$ and $PM_{10}$ traces. As the results from this correlation analysis show, reasonably strong correlations may be observed between mass and number concentration time series for individual journeys (or parts of journeys), but no general conclusion can be drawn due to the high variability of the correlations observed for a large number of journeys.

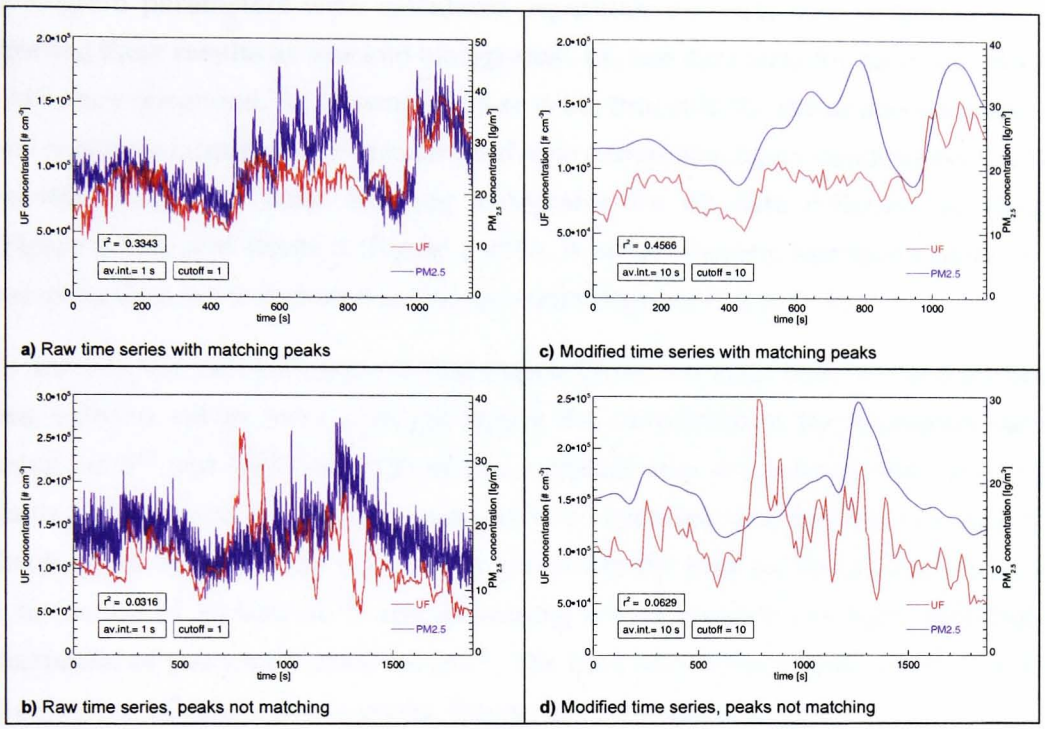


**Figure 4-9 Correlation between ultrafine and $PM_{2.5}$ particles, example plots illustrating the reasons for variability in $r^2$**

This further supports the hypothesis that $PM_{10}$ and $PM_{2.5}$ mass concentrations behave similarly, but they both differ from ultrafine particle concentrations and can thus not be used as a surrogate for in-vehicle particle concentrations in general.

## 4.6 Stacked histograms

While time series plots allow the visual inspection of raw data in terms of concentration behaviour over time (Section 4.3), histograms give an assessment of how the concentration measurements are distributed between minimum and maximum values by grouping them into bins. In order to aid the inspection of multiple histograms, a visualisation method was developed similar to the 'data strip' method used for the visualisation of time series data in Section 4.3. This technique was used to inspect the field study time series data to investigate the distribution of the measured values and to identify any underlying pattern.

### 4.6.1 Methods

Histogram parameters were calculated separately for each data collection run. By plotting these results as 'stacked histograms', i.e. one data strip for each run, synoptic plots were generated. As shown in Figure 4-10, this aids the simultaneous inspection of a relatively large number of individual time series. The figure contains two plots for evening rush hour data, showing histograms for all data collected on Route 1 (Figure 4-10a) and Route 2 (Figure 4-10b). A set of synoptic histogram plots for the complete data set is included in the appendix (Figures B-2 to B-4).

In order to use this technique to investigate where the main body of the distribution lies, extreme values were removed before the calculation of the histogram data by using the $5^{th}$ and $95^{th}$ percentile of the complete data set as lower and upper cutoff limits, respectively. Each circuit completed is represented in the plot as a data strip, divided into 30 bins. Low values are recorded in the bins on the left hand side, with a minimum of 16,000 $cm^{-3}$, and increasing values towards the right, reaching the maximum of just under 250,000 $cm^{-3}$. The bins have been colour-coded to indicate the number of values they contain. Empty bins are shown in grey.

### 4.6.2 Results

Comparison of routes

Typically, the data collected during a journey do not cover all bins, i.e. either some of the upper or lower bins remain empty. Although there is a lot of variability in the data, it is clear that most values fall in the lower bins while peak concentrations occur comparatively infrequently.

The two plots in Figure 4-10 also demonstrate how stacked histogram plots can facilitate the comparison of different data sets. For example, the data for Route 1 tend to accumulate in the lower bins, generally staying below 100,000 $cm^{-3}$, whereas data from Route 2 are more likely to exceed this value. Moreover, Route 2 data seem to be more variable, and stretched over many bins (immediately apparent as longer coloured data strips). On this route, peak values are generally more often recorded than on the other routes. These observations are in accordance with the findings from Section 4.2 where average concentrations were shown to be generally higher on Route 2. A comparison between plots for morning and evening data (Appendix B.3) shows that the data strips for morning data tend to be slightly shifted towards the right end

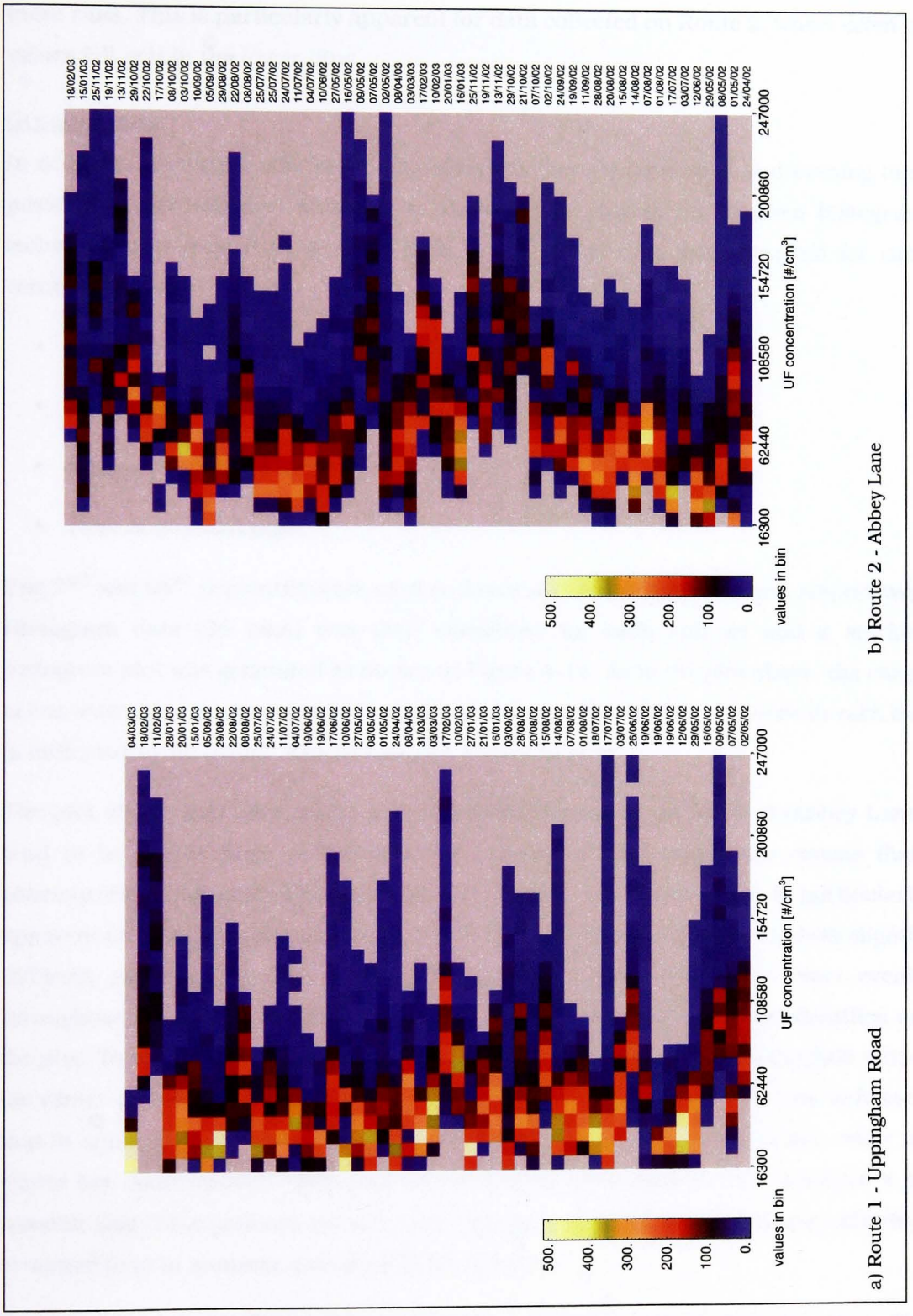


**Figure 4-10 Stacked histograms for data runs carried out on Routes 1 and 2 during evening rush hour**

of the scale, indicating that measured concentration values are typically higher on these runs. This is particularly apparent for data collected on Route 2, where often no values fall within the lower bins.

Grouped data

In order to investigate whether such differences between morning and evening (and possibly winter/summer) data can be shown more clearly, the 'stacked histogram' technique was applied to grouped data. A sub-set of data was compiled for each combination of:

- Location (Routes 1 to 3).
- Direction of travel (clockwise/anti-clockwise).
- Season (summer/winter).
- Time of day (am/pm).

The $2^{nd}$ and $98^{th}$ percentile were used as lower and upper cutoff values, respectively. Histogram data (30 bins) was then calculated for each sub-set and a stacked histogram plot was generated as shown in Figure 4-11. As in the plot above, the range of concentration values is marked on the x-axis and the number of values in each bin is indicated by its colour. Empty bins are shown in grey.

The plot shows that once again, concentrations measured on Route 2 (Abbey Lane) tend to be higher than on the other two routes. Closer inspection reveals that, concentrations are generally higher in winter than in summer. This is particularly apparent for data from Route 2 and Route 3. The histograms for Route 1 show slightly different patterns. On this route winter data is clearly distributed more evenly throughout the bins than summer data, i.e. no clear maximum can be identified on the plot. To some extent this tendency can also be observed for most of the data strips for winter data from the other routes, although it is less pronounced. This indicates that in summer lower concentrations are more consistent, with few peaks, while in winter low concentration values are as likely to occur as high values. However, it is possible that these patterns are not truly representative, since less data was collected in winter than in summer, particularly for Route 1.

Figure 4-11 also confirms the observation described in Section 4.2 that the concentration values from morning runs are generally higher than those from evening

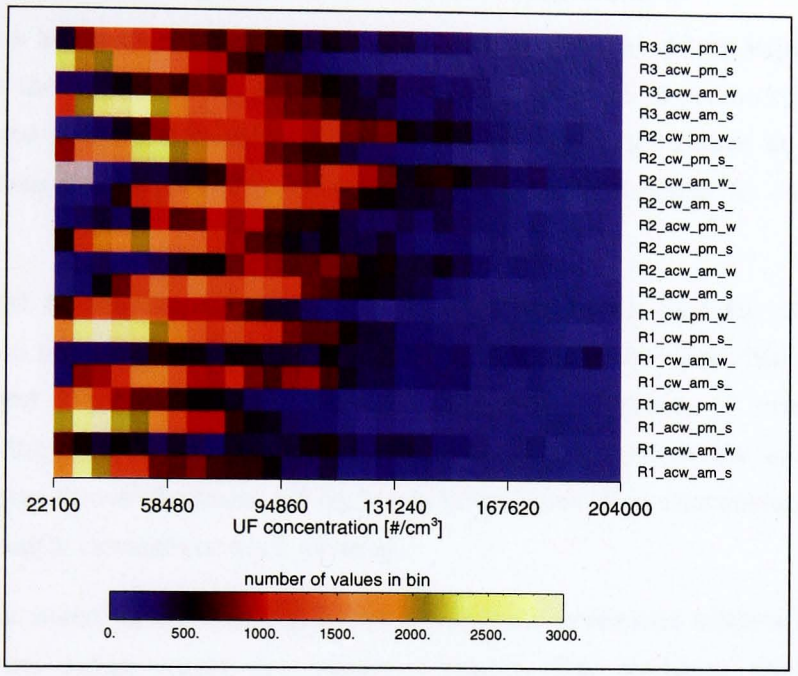


**Figure 4-11 Histograms for grouped data: ultrafine particle concentrations measured on field study routes**

*R1 - Uppingham Road, R2 - Abbey Lane, R3 - City Centre*

runs. Again, this is particularly apparent for Route 2 data but is also noticeable for the other routes. As discussed in Section 4.2, these effects are likely to be due to a combination of factors affecting particle concentrations, namely meteorological conditions limiting dispersion in winter and in the mornings, and cold start emissions, which are higher at lower temperatures, i.e. during winter and morning runs. On routes with higher traffic flow, such as Route 2, these effects are likely to be particularly pronounced. However, the noticeable differences between the histograms for Route 1 and those for the other two routes are thought to be due to less field study data being available for this route. For details about the individual road links refer to Table 3-3.

### 4.6.3 Summary

The stacked histogram technique was used to inspect the distribution of measurements from multiple data collection runs, leading to the detection of underlying patterns, which further consolidate the findings from Section 4.2, indicating that meteorological and location specific parameters may cause much of the observed variations in particle concentrations.

## 4.7 Influence of peak values on exposure

The previous sections have provided evidence that elevated driver exposure levels result from the influences of numerous parameters, both in terms of short-term variability and long-term averages. The aim of the analysis presented in this section was to investigate the contribution of peak concentration levels to overall driver exposure.

As discussed in Section 2.3, personal exposure is determined by the pollutant concentration and the duration of the exposure. In order to compare driver exposure from different journeys, average exposure values (i.e. average of concentrations throughout the journey) are typically used. However, the time series and histogram plots presented above (Sections 4.3 and 4.6) have shown that concentrations tend to vary significantly throughout each journey.

Although, as noted in Section 2.2.1, the underlying processes influencing particle concentrations along roads are rather complex, the resulting effects on the concentration levels experienced by road users can be generalized using a basic hypothesis, as is illustrated in Figure 4-12. Based on minimum assumptions and findings from the preliminary analysis, it seems reasonable to assume that an effective background of ultrafine particle concentrations exists in the traffic stream, and that the observed concentration peaks are due to distinct pollution events which are superimposed on this background. These pollution events are thought to be caused by pockets of high concentrations, resulting from transient events, such as high emissions from the vehicle in front and/or reduced dispersion of pollutants in certain areas, e.g. in street canyons. When a vehicle enters such as pollution pocket, in-vehicle concentrations are likely to rise also.

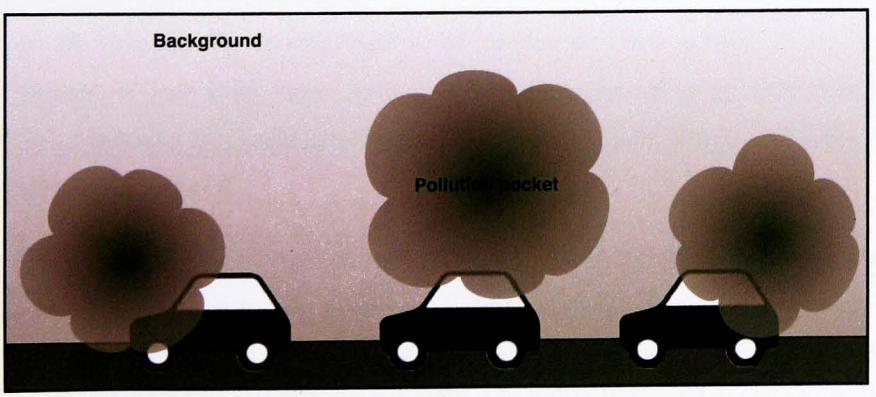


**Figure 4-12 Hypothesised pollution scenario**

As discussed in Section 2.3.2, it is not clear how exactly vehicle speed, ventilation settings and other car specific parameters affect the ingress of particulate matter into the passenger compartment. However, the investigation of time series traces in Section 4.3 has shown that concentration peaks can be associated with distinct events, such as, for example, following strongly polluting vehicles and queuing at traffic lights. These observations thus support this hypothesis. The apparent background in the pollution corridor may in fact be a result of pockets of high concentrations of ultrafine particles that have been mixed and partially dispersed by vehicle induced turbulence.

In order therefore to investigate the influence of these transient events on driver exposure levels, a method is proposed here which aims to separate distinct concentration peaks from the 'background'. This allows an investigation of the proportion of total driver exposure that be attributed to elevated concentrations, and thus helps to gain an understanding of the temporal dynamics of driver exposure. Moreover, the results may also have implications for health studies. As outlined in Section 2.2.3, the underlying biological mechanisms for the association between particulate matter and its effects on health have not yet been clearly identified. However, although exact limit values are difficult to establish, elevated particle concentrations have been linked to increased health effects, and repeated high short-term exposure episodes are thought to contribute to chronic effects (e.g. Weijers et al., 2001).

The following sections outline the details of the proposed method and discuss its merits and shortcomings alongside the derived results.

### 4.7.1 Setting the threshold

In order to determine the contribution of elevated concentrations to the overall exposure values, it was important to first established an appropriate definition of a 'peak'. Various options for setting a threshold, values above which would be regarded as peaks, were considered.

Ideally, a limit value, concentrations below which were deemed to cause no adverse health effects, should have been chosen as a threshold. However, there is currently no such limit value available for concentrations of ultrafine particles. Moreover, it is likely that such a limit would be based on average concentrations for a relatively long time

interval of 15 minutes or 1 hour, whereas this study has used high frequency data (i.e. 1 Hz) in order to investigate the short-term temporal dynamics of exposure.

Another option was to use a 'reasonable' percentile value to set the threshold, e.g. all values above the $95^{th}$ percentile for a certain time series would be considered peak values. It was first supposed to calculate this percentile value based on the complete data set and apply it as a universal threshold to all run time series. However, visual inspection of a number of time series showed that this would in many cases have resulted in entire runs being classified as 'peak' or 'non-peak', i.e. all values in the time series above or below the universal threshold, respectively. With this method it would therefore not have been possible to identify the influence of discrete events, since 'peaks' were found to contribute either 0 or 100 % to overall exposure values on many runs. This option was therefore disregarded, as was the alternative option of using run-specific thresholds, since it the selection of a percentile threshold setting was found to be too arbitrary to provide meaningful results.

Following these considerations, it was decided that the most useful results could be derived by using a parameter representing background concentrations. This was based on the hypothesis of particle concentrations consisting of emissions from local sources in addition to background concentrations. However, it is difficult to separate background concentrations and additional contributions. Ideally, measured background concentration data should have be used. However, background particle concentration data from monitoring stations in Leicester only include $PM_{10}$ measurements, not $PM_{2.5}$ or ultrafine particle data. Moreover, due to the high spatial variability of the smaller particles, it is unlikely that an UF background concentration as such exists on a city or area wide scale.

It was therefore decided to introduce a 'background substitute' (BS) parameter which was derived separately for each run, based on the lowest value measured. It is accepted that this value is not the true background concentration, since local sources will have contributed to it to some extent. However, it is thought to represent an approximation of a corridor (in-traffic) specific minimum, values above which are additionally elevated.

Figure 4-13 shows two time series of 10 minutes duration, one with a generally higher values (blue) than the other (red). The raw time series data (black) have been smoothed to reduce the short-term variability in order to be able to identify peaks

more easily. A dashed line is included in the plot for each time series, indicating the BS threshold. The area between time series trace and the associated dashed line represents the contribution of elevated concentrations to the total exposure experienced during the journey. The ratio between this value and the total journey exposure gives an indication of how strongly the overall exposure is affected by concentrations exceeding the BS threshold value.

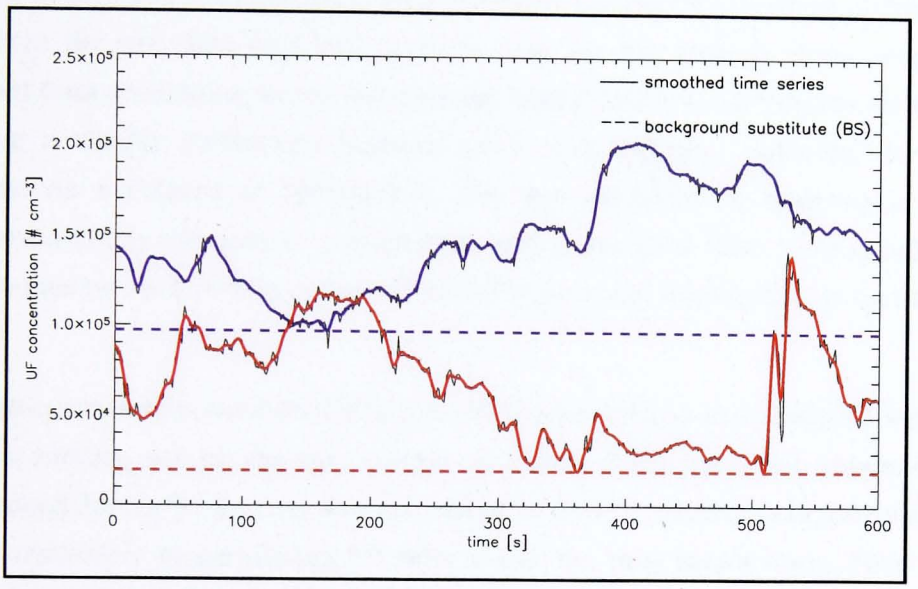


**Figure 4-13 Example of two time series plots showing contribution of peak values to overall exposure (journey 1 blue, journey 2 red)**

The results for the two example time series, shown in Table 4-5, illustrate how this ratio can vary between different journey. Although the total contribution of elevated concentrations to overall exposure is similar for both journeys (3.06 x$10^7$ $cm^{-3}$ s and 2.83 x$10^7$ $cm^{-3}$ s, respectively) their proportional contribution clearly varies. The BS value for journey 2 (red in Figure 4-13) is comparatively low, resulting in a peak contribution of over 70 %, while journey 1 (blue in Figure 4-13) is dominated by the high BS concentrations, with peaks only contributing 34 % to overall journey exposure.

| | total expos. (TE) [$cm^{-3}$ s] | 'background substitute' (BS) [$cm^{-3}$] | total contribution of elevated conc. (PE) [$cm^{-3}$ s] | proportional contribution of elevated conc. (FRAC) [%] |
|---|---|---|---|---|
| **journey 1** | 8.88 x$10^7$ | 9.69 x$10^4$ | 3.06 x$10^7$ | 34 |
| **journey 2** | 3.99 x$10^7$ | 1.94 x$10^4$ | 2.83 x$10^7$ | 71 |

**Table 4-5 Numerical comparison of exposure composition for two time series**

In order to investigate how these exposure parameters vary throughout the whole data set, all time series from the field campaign were analysed by calculating BS values and exposure ratios and subjecting the results to a statistical analysis. The methods and results are explained in the following sections.

### 4.7.2 Data preparation and calculations

Custom written IDL routines were used to extract the relevant sections of time series data from the raw data files and prepare them for the analysis. This preparation included data smoothing which reduced the high frequency oscillations in the data, allowing a clearer distinction between peak and non-peak episodes during the analysis. As explained in Section 4.3, this was achieved by applying a Fourier transformation to the data in combination with a low pass filter. This results in all periodicities below a certain cut-off value being removed, leading to a 'smoother' time series.

Each time series was smoothed with the described method and analysed separately using a further set of custom written routines. First the total exposure (TE) experienced during the journey was calculated as the sum of all data in the time series. This is equivalent to calculating the area under the time series trace. Then the BS threshold value was determined, using the lowest value of the smoothed time series trace (dashed line in Figure 4-13). By calculating the area under that line and substracting the result from the total exposure, a 'peak exposure' value (PE) was derived, i.e. the portion of the total journey exposure which can be attributed to elevated concentrations. By calculating the ratio between peak and total exposure, a value for the fraction of total exposure attributable to elevated concentrations (FRAC) was derived. This values thus gives an indication as to what extent the experienced exposure is associated with discrete pollution events in addition to the general concentrations in the traffic corridor.

The results for these parameters were collated for all time series (N = 307) and subjected to a statistical analysis using the EViews statistical package. Descriptive statistics and equality tests of means were calculated in order to investigate whether the results varied significantly depending on where or when the exposure data was collected.

### 4.7.3 Results

The descriptive statistics listed in Table 4-6 show that both total exposure and BS threshold vary strongly between different journeys, i.e. three and one orders of magnitude, respectively. The fraction of total exposure contributed by elevated concentrations varies from 17 % to 79 %, indicating that during certain journeys, concentrations are only slightly elevated above the background substitute value, while on other journeys, total exposure is strongly influenced by high concentrations.

| | TE [$cm^{-3}$ s] | BS [$cm^{-3}$] | FRAC [%] |
|---|---|---|---|
| **mean** | $8.53 \times 10^7$ | $4.21 \times 10^4$ | 48 |
| **median** | $7.94 \times 10^7$ | $3.66 \times 10^4$ | 47 |
| **maximum** | $2.29 \times 10^8$ | $1.25 \times 10^5$ | 79 |
| **minimum** | $7.84 \times 10^6$ | $1.13 \times 10^4$ | 17 |
| **standard deviation** | $4.68 \times 10^7$ | $2.18 \times 10^4$ | 12 |

**Table 4-6 Descriptive statistics for journey exposure composition (N = 307)**

As the scatterplot in Figure 4-14 shows, the contribution of elevated concentrations to overall exposure tends to be lower for high BS threshold values ($r^2 = -0.27$, significant at $p < 0.0001$). This indicates that during periods of generally high concentrations, exposure tends to be dominated by the 'background' and peak values contribute only in small proportions. The correlation between BS and FRAC, however, is rather weak, with particularly pronounced scattering at lower values. It is possible that these results are biased by the use of the journey minimum as an indicator for background (BS), i.e. during longer journeys, the chance of recording more representative minimum concentration values would have been higher, which may have led to a clearer relationship between FRAC and BS values.

Equality tests of the mean were carried out for BS and FRAC values from all journeys in order to investigate whether the variation in those parameters can be associated with journey specific parameters such as when and where they took place.

A significant difference ($p < 0.0001$) was found between 'background substitute' (BS) values from summer and winter journeys as well as between morning and evening journeys, with generally higher values recorded during summer and morning journeys. A less significant difference ($p = 0.0038$) was found between the contribution of elevated concentrations to overall exposure (FRAC) during morning

and evening journeys, with higher values occurring in the evening, while no statistical evidence was found for an effect of seasonal variation on FRAC ($p = 0.25$).

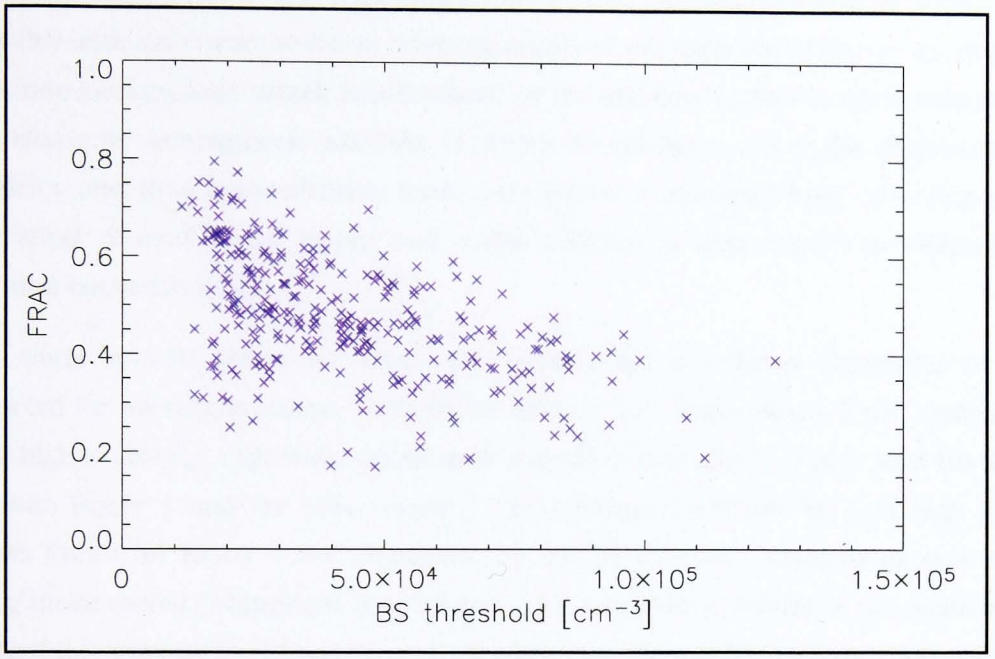


**Figure 4-14 Scatterplot for BS threshold vs. FRAC ($r^2 = 0.27$, significant at $p < 0.0001$)**

There is also evidence that both parameters are influenced by the route taken. Significant differences were found between FRAC and BS results from journeys on Route 1 compared to the other two routes (all $p < 0.001$). The route was associated with generally lower values of FRAC and higher values of BS compared to results for Routes 2 and 3.

### 4.7.4 Discussion

The findings regarding the 'background substitute' value (BS) generally follow the observations made for the seasonal and diurnal variation in exposure values, i.e. higher values are associated with morning and winter periods as opposed to evening and summer periods. These variations are thought to be due to the influence of meteorological parameters, as discussed in detail in Section 4.2. Based on the strong correlation between BS and total journey exposure TE ($r^2 = 0.45$, $p < 0.0001$) it is very likely that the results are due to the same effects and thus the same reasoning applies. The results for the fraction of total exposure attributable to elevated concentrations (FRAC) show an inverse diurnal variation, which is likely to be an

effect of its negative correlation with BS. It is not clear why there is a lack of evidence for seasonal variation.

These findings support the hypothesis that a road can be regarded as a pollution corridor with somewhat uniform concentrations of ultrafine particles, i.e. an effective pollution background, which is influenced by the ambient meteorological conditions. For example, atmospheric stability is likely to similarly affect the dispersion of particles and thus concentration levels throughout a relatively large area (e.g. city). The effect of local street layout and traffic conditions then results in additionally elevated concentrations.

The route specific difference found for both FRAC and BS is dissimilar to that reported for average exposure reported in Section 4.2. While Route 2 was associated with higher average exposure values than the other two routes, FRAC and BS differ between Route 1 and the other routes. The comparatively low BS and high FRAC values found for Route 1 are thought to be due to exposure patterns on this route being more varied throughout the journey. This could be a results of the route being divided into two parts in regards to road types, i.e. the northeastern section consists of main roads while the other links are smaller local roads. It is therefore likely that low concentrations are recorded on the less polluted local road links (giving low BS values) and elevated concentrations strongly exceeding these values on the busy main roads (leading to high FRAC values). Routes 2 and 3 are more consistent in terms of road types, which may explain why no significant difference was found between the results for those routes.

In conclusion, the parameter FRAC proposed here aided the investigation of the contribution of elevated concentration levels to overall journey exposure. The results show that the proportions contributed to total measured exposure from peak and background concentrations vary significantly between different journeys, which can be partly attributed to the influence of route and meteorological parameters. However, although the analysis has yielded some plausible associations, the findings also indicate that FRAC may be biased due to the 'background substitute' parameter which is used to calculate it. If actual measured background values were used instead, the proposed analysis method may lead to more meaningful results. When such data are not available, general descriptive statistics based on average journey exposure values are likely to be as suitable for the analysis of personal exposure.

## 4.8 Summary

The descriptive statistics calculated for the field study data set indicate that meteorological parameters affect in-vehicle concentrations of ultrafine particles as well as particles in the $PM_{10}$ and $PM_{2.5}$ ranges. Higher values of all particles were measured during morning runs. For ultrafine particles, winter runs were also associated with elevated in-vehicle concentrations, while mass concentration values (i.e. $PM_{10}$ and $PM_{2.5}$) tended to be higher during the summer months, indicating that mass and number concentrations react differently to variations in meteorological conditions. The data also suggested that all measured particle concentrations were affected by 'route', i.e. road specific factors. Time-series plots and histograms of raw data indicated that the observed effects are also apparent when inspecting high resolution measurement data from multiple runs. An analysis of the contribution of elevated concentrations to overall UF exposure further confirmed the effects of meteorological and route specific factors on exposure values.

The inspection of high resolution time-series plots showed that ultrafine particle concentrations exhibit less short-term variability than $PM_{10}$ and $PM_{2.5}$ concentrations. The plots further suggest that traffic events affect in-vehicle concentrations, such as following a particularly polluting vehicle or experiencing periods of queuing or acceleration. However, the particle concentrations of the different size ranges seem to vary strongly in their sensitivity to those events, i.e. many instances of elevated ultrafine concentrations are not mirrored in the $PM_{10}$ and $PM_{2.5}$ time-series.

The correlation analysis showed that this has a strong effect on the correlation between the different size ranges. Results indicate that $PM_{10}$ and $PM_{2.5}$ concentrations are relatively closely correlated, particularly for longer averaging intervals, while there is no consistent correlation between the time-series of mass concentration measurements and those of ultrafine particle number concentrations.

Results from the investigation of concentration peaks indicate that a large proportion of driver exposure can be attributed to concentrations resulting from transient traffic events which are elevated above an apparent background level that may vary considerable from day to day and between different locations.

The results show that, generally, particle concentrations inside the vehicle follow the patterns found in other studies for outside concentrations. Most importantly however, the findings support the hypothesis that mass concentrations of the larger particles cannot be used as surrogates for number concentrations of ultrafine particles. In order to fully investigate driver exposure it is thus necessary to take all size fractions into account by measuring them separately.

Chapter

# 5 *Identifying the Main Determinants*

As discussed in the previous chapter, driver exposure to particulate matter seems to be influenced by a combination of external factors such as seasonal variations in meteorological conditions and route specific parameters. A multiple regression analysis was carried out in order to formally investigate these apparent effects. The analysis aimed to identify which parameters (determinants) most affect in-vehicle concentrations and driver exposure by applying standard statistical methods to the field study data set. The analysis focusses on ultrafine particles but $PM_{10}$ and $PM_{2.5}$ data are considered also for comparative purposes.

## 5.1 Introduction

One commonly used method of identifying the main determinants affecting concentration or exposure values is multiple linear regression analysis. The use of this technique is based on the assumption that the relationships between personal exposure and the predictor variables are linear. Even though this may simplify the possibly complex underlying relationships, using linear equations requires less complex computations (i.e. faster processing) and it has been shown that linear equations often provide a good approximation even if the true relationship in not linear (Allison, 1999). Moreover, it is relatively easy to modify the linear equations for certain cases of non-linearity, such as skewed data and categorical variables. This is described below.

Multiple linear regression is regularly used in air quality studies as a tool to investigate the complex underlying influences of parameters affecting pollutant concentrations. Parameters investigated range from meteorological conditions (e.g. temperature, wind speed) to road related variables (e.g. road width, speed limit) and types of activity (e.g. type of road transport used). For example, Adams et al. (2001b) used multiple linear regression to identify which type of road transport could be associated with the highest exposure values and to investigate which other parameters could be shown to have an effect on exposure levels. A study by Johnson and Ferreira (2001) used multiple regression analysis to develop a model to predict road emissions based on measured parameters such as meteorological data, traffic flow and speed as well as vehicle fleet composition. Although the study focussed on the prediction of emissions from traffic flows, the methodology clearly shows strong similarities to the analysis carried out by Adams et al. (2001b). Multiple linear regression analysis is evidently considered a useful tool in air quality research. Although other methods are available, multiple regression analysis is deemed to strike a good balance between complexity and accuracy. For example, Chaloulakou et al. (2003) compared predictions of daily $PM_{10}$ averages derived with various different models, one being multiple regression and the other neural network analysis. The predictor variables were selected by using a stepwise regression analysis procedure and the final regression models included meteorological parameters and time-lagged $PM_{10}$ concentration measurements. The fact that stepwise multiple regression was chosen in a model comparison study to select input parameters for different types of $PM_{10}$ model, confirms that this technique is a commonly used method in the analysis of air quality data.

Like most statistical methods, multiple linear regression analysis can only be applied if the data set meets certain criteria. The development of an empirical regression model therefore comprises data preparation, selection of the correct statistical methods and a comprehensive test procedure for testing and interpreting the results. It is an iterative process, which is repeated until the 'best' model is found, i.e. the model with the least input parameters still providing a 'good' fit.

The two main objectives of the multiple regression analysis carried out here were: (1) to identify the main determinants of driver exposure; and, (2) to evaluate the potential use of driver exposure field data for long-term and journey exposure modelling. The

analysis carried out to achieve the first objective is described in the following sections, together with general descriptions of the statistical methods used. The results of this stage of the analysis then formed the basis for work on the second objective, which is the focus of Chapter 6. The methods and variables described below were chosen to be suitable for both phases of the analysis.

## 5.2 Theoretical background of multiple regression analysis

This section gives an overview of the aspects of multiple regression analysis which are most relevant for the identification of the main determinants of driver exposure, such as variable types and assessment of the accuracy of derived coefficients and models. Further details for some of the issues are included in Appendix C.1, indicated by cross-references where applicable.

The general purpose of multiple linear regression is to learn about the relationships between a number of independent or predictor variables and one dependent variable. The computational problem that needs to be solved is to fit a straight line to a number of points. The simplest case of one independent and one dependent variable can be visualized in a scatterplot. For $p$ independent variables this line would lie in an $p$-dimensional space and can be described as

$$y = b_0 + b_1x_1 + b_2x_2 + ... + b_px_p \quad \textbf{(5-1)}$$

where $x_1$, $x_2$,..., $x_p$ are the independent variables and $b_1$, $b_2$,..., $b_p$ are a set of scalars. This can be extended to a whole set of observations as shown in Eq 5-2 (StatSoft, 2004). If there are $n$ observed values of $Y$ and $n$ associated observed values for each of $p$ different $X$ variables then $Y_i$, $X_{ip}$, and $e_i$ can represent the $i$th observation of the $Y$ variable, the $i$th observation of each of the $X$ variables, and the $i^{th}$ unknown residual value, respectively.

$$Y = \begin{bmatrix} Y_1 \\ \dots \\ Y_n \end{bmatrix}, X = \begin{bmatrix} 1 & X_{11} & \dots & X_{1p} \\ \dots & \dots & \dots & \dots \\ 1 & X_{n1} & \dots & X_{np} \end{bmatrix}, e = \begin{bmatrix} e_1 \\ \dots \\ e_n \end{bmatrix} \quad \textbf{(5-2)}$$

In matrix form, i.e. looking at the complete vector for the dependent variable $Y$ and the corresponding matrix of values for the independent variables $X$ this can be written more simply as

$$Y = Xb + e \tag{5-3}$$

where $b$ is a column vector of 1 (for the intercept) and $p$ unknown regression coefficients and e the vector of unknown residual values. The matrix form is particularly useful for an efficient design of custom written statistical analysis routines.

### 5.2.1 *X* and *Y* variables

$Y$ is called the 'dependent variable' and $X$ is referred to as the 'independent variables' or 'predictor variables'.

Independent variables can be categorical or continuous. A categorical variable is a variable measured on a nominal scale, i.e. it takes one of a set of discrete values (for instance 'type of road'). In order to use such a variable in a multiple regression analysis, a separate predictor variable is coded for each category identified by the variable. This is achieved by adding as many 'dummy' variables to the matrix as there are categories and assigning the value of 1 to the variable representing the category that applies and 0 to all the others. However, including all dummy variables in the analysis would lead to 'extreme multicollinearity', i.e. it would be impossible to get separate estimates for the coefficients for those variables. To avoid this problem, one of the dummy variables is chosen as the reference category and excluded from the regression analysis. Each of the derived coefficients for the remaining categories therefore represents a comparison between an included category and the reference category (Allison, 1999). If a variable can only take one of two states, it gets coded as one dummy column. To further specify variable types, categorical variables whose categories represent qualitative divisions but are continuous (e.g. 1 - strongly disagree, 2 - disagree, 3 - agree, 4 - strongly agree) are called 'categorical ordinal' as opposed to 'categorical nominal' where categories do not follow any order (Everitt and Dunn, 2001).

It is recommended to inspect scatter plots of the potential predictor variables prior to the multiple regression analysis to investigate the relationships between those

independent variables and between the dependent and independent variables. This helps to gain a better understanding of the composition of the data set and to identify correlations (linear or otherwise) between variables. Scatter plots may also reveal problems with near-extreme multicollinearity between two or more of the independent variables, which makes it difficult to reliably estimate the coefficients for those variables. For further details on implications and solutions for this problem refer to Appendix C.1.1.

### 5.2.2 Assessing the validity of a regression model

The derived coefficients $b$ form the basis of a model that allows the calculation of predicted values for the dependent variable from the independent variables. To assess the 'goodness of the fit' of a regression model, a multiple correlation coefficient $R^2$ is calculated. This takes values between 0 and 1, with values close to 1 indicating a good fit. The correlation coefficient also indicates which proportion of the variability in the dependent variable is explained by the model.

In order for the described statistical methods to provide reliable results, it is desirable that the residuals of the model are normally distributed. This can be aided by making sure that the dependent and independent variables are approximately normally distributed. Air quality data such as exposure values tend to be strongly positively skewed (Barratt, 2001). A common approach is to logarithmically transform such data to obtain a normally distributed data set. If such a transformation is applied to the dependent variable of a multiple linear regression, the relationship between $X$ and $Y$ as shown in Eq 5-3 is modified as follows

$$Y = e^{Xb+e}. \tag{5-4}$$

Descriptive statistics for such variables ($Y$) are therefore based on the geometric mean and standard deviation rather than the commonly used arithmetic mean and standard deviation.

The normality of the variables and the residuals is easily assessed by inspecting histogram plots, probability plots and using the Jarque-Bera hypothesis test for normality. Testing for normality is particularly important for small samples. For larger samples, statistics such as confidence intervals and hypotheses tests will be

good approximations even if the residuals are not normally distributed (Barratt, 2001; Allison, 1999). For more details on these methods refer to Appendix C.1.3.

Another assumption for the use of the described statistical methods is homoskedasticity (equal variances), i.e. that the degree of scatter around the regression line is roughly the same for all $x$-values. This is best assessed by inspecting an equal variance plot of the data, which is a scatterplot of the unstandardized residuals of the model vs. the predicted values for the dependent variable. Or, alternatively, White's test for homoskedasticity can be used, which provides a probability for the null hypothesis of 'homoskedasticity' being valid (i.e. small probability - greater chance of heteroskedasticity). For further details refer to Appendix C.1.3. If data exhibit heteroskedasticity, the calculated standard errors can be adjusted by using White's method for calculating robust errors which is available as a function in the EViews statistical package (QMS, 2002).

### 5.2.3 Assessing the accuracy of the regression coefficients

Since the derived regression coefficients $b$ are calculated based on a sample of measured values, they only represent estimated values for the actual relationship between the variables. It is therefore necessary to assess their accuracy by calculating confidence intervals and/or using hypothesis tests. In this study, hypothesis test are used to determine (1) whether the derived coefficients are significantly different from zero (t-statistic) and (2) whether the predictor variables contribute significantly to the model (F-statistic). For further details on the methods used refer to Appendix C.1.4.

Sample size has a profound effect on tests of statistical significance. Since statistical tests are only approximations, they can be reasonably accurate for large samples (e.g. $n > 1000$) but are not always reliable when small samples are used. However, for large samples, regression coefficients can be significant even if no true relationship exists. This means that the magnitude of the coefficients has to be checked to make sure that only truly influential variables are considered in the analysis (Allison, 1999).

### 5.2.4 Backward elimination

The method of using F-tests for conditional significance testing could be used to try all combinations of independent variables and to select the most suitable, usually the model with fewest variables that is compatible with the data. Since there are $2^p$-1

possible models that would have to be investigated, other methods of finding the 'best' model are used (Everitt and Dunn, 2001). One commonly used method is backward elimination, which starts with the most complicated of the possible models, i.e. all variables included. An F-statistic and p-value are calculated for the exclusion of each variable from the model in turn and the least significant variable (highest p-value) is then removed from the model, provided that the significance level is above some predetermined level. In cases where many variables are significant (p~0) the next variable removed is the one whose removal results in the least reduction of the multiple correlation coefficient $R^2$. This process is repeated until the simplest model compatible with the data is found. Finding a model with a good fit that only includes a few variables will ease the potentially difficult problem of interpreting the results (Everitt and Dunn, 2001).

In order to estimate which of the variables included in the final model is the most influential, i.e. explains most of the variability in the dependent variable, the ratio of the between groups sum of squares to the total sum of squares ($R^{2*}$) is calculated for each independent variable. This $R^{2*}$ takes a value between 0 and 1, with higher values occurring for more influential variables. For example, if $R_{xi}^{2*} = 0.3$, the variable $x_i$ explains 30 % of the variability in the dependent variable.

## 5.3 Methods and data used

A statistical analysis was performed in order to identify the main determinants (i.e. predictor variables) of driver exposure. A set of IDL and EViews routines were custom written to facilitate the extraction of relevant data from the main data set, to inspect the selected data prior to the analysis and to carry out the calculations required for model development and assessment, including inspection of residuals.

The analysis utilizes a data set that is based on the main data array described in Section 3.2.4. The data set contains meteorological and link specific parameters as well as various aggregated variables derived from particle concentration measurements, such as average concentration and average total exposure. The data were extracted from the IDL data array as records, each record containing the dependent variable investigated (i.e. link averages for concentration, total exposure or total exposure/km) and values for each of the potential predictor variables. These

records are combined in a two-dimensional array which is fed into the IDL routines that run through the steps of the multiple regression analysis as described below.

The data set was prepared for the analysis by selecting all complete records from the main data array (Section 3.2.4), i.e. only where field data as well as values for all other included variables were available. Although techniques for replacing missing values exist, they would have introduced synthetic (i.e. interpolated) data, and were thus not used in this analysis.

The data set was analysed for suitability and then subjected to a multiple regression analysis with backward elimination. The origin and nature of the dependent and independent variables used are outlined in the following sections.

### 5.3.1 Dependent variables

As noted above, the results from this analysis were to form the basis for the modelling of journey exposure in the next chapter. Therefore, the three dependent variables for this multiple linear regression analysis were selected based on the requirements of the modelling scenarios described in Section 6.1.1.

First a model equation was derived for 'average total exposure per link' (ATE). The values were derived by calculating the sum of all measured concentration values collected on a particular link for each visit and then taking the average of all visits during one journey. Since a link would typically be visited two or three times during one journey, the resulting 'average total exposure' calculated from all visits on one run is more representative of conditions on that link than the average value for an individual visit, which may be exceptionally high if an extreme pollution episode occurred during that visit. These values for the dependent variable were calculated during the data processing phase and stored in the main data array (see Section 3.2).

In order to investigate total link exposure independent of link length, 'total exposure per link per km' (ATEK) was chosen as an alternative dependent variable in the potential driver exposure model. This was calculated by dividing each 'total exposure per link' value by the link length.

The third dependent variable investigated was 'average concentration' (AC), which was derived by calculating the mean of all values measured on a link during one journey (i.e. during consecutive visits). The geometric, rather than the arithmetic, mean was

used due to the strongly skewed distribution of the data (i.e. normally distributed after logarithmic transformation).

All three dependent variables were logarithmically transformed to achieve a distribution more closely resembling the normal distribution. After the exclusion of outliers for the dependent variable, i.e. values exceeding the $95^{th}$ percentile limit, and excluding all observations with extreme values for any of the independent variables (explained below), the dependent variable for the ATE case was clearly normally distributed (Jarque-Bera $p = 0.97$). Although the this test was less conclusive for the dependent variables from the other two cases ($p=0.00005$ for ATEK and $p=0.0002$ for AC), normal probability plots produced for these data showed nearly straight lines. Due to this fact and the large number of observations considered, it was decided that the data was sufficiently normally distributed to be used in the analysis.

### 5.3.2 Independent variables

Table 5-1 gives an overview of the 19 independent variables originally considered as potential predictor variables. They include physical link parameters (e.g. link length, road function) as well as meteorological parameters (e.g. temperature, wind speed). Physical parameters for each link were collected as part of the field study, apart from link length which was derived from GIS maps.

The variable 'URR' (usable road ratio) was derived from the number of lanes on a link minus the space taken up by vehicles parking at the road side (i.e. 1/2 lane, 1 lane).

The variables 'road function' (RF) and 'link end description' (LED) are categorical variables with five categories each (link function: local, high street, inter-urban, arterial/in, arterial/out, link end description: open, left turn, signal, right signal, right turn without signal). The variable 'crossing' (C) denotes the existence of one or more pedestrian crossings on the link (categories: yes and no) and the variable 'road side environment' (RE) indicates whether the link is located in a residential or non-residential area. A description of these parameters and categories is given in Section 3.1.3.

Meteorological data, provided by the Pollution Control Group of Leicester City Council, consisted of hourly averages for temperature (at 2 m), temperature difference (between 2 m and 10 m), wind speed, wind direction, maximum wind

speed, global radiation and rainfall. Since data collection runs were carried out during morning and evening rush hours, i.e. approx. 8 - 9 am and 5 - 6 pm respectively, the hourly average value for the $9^{th}$ hour and the $18^{th}$ hour in each day, respectively, were used in the analysis.

| Variable number | Variable name | Abbreviation | Variable type | Unit |
|---|---|---|---|---|
| 1 | link length | LL | continuous | km |
| 2 | TRIPS flow | tFL | continuous | #/h |
| 3 | TRIPS speed | tSP | continuous | mph |
| 4 | usable road ratio* | URR | continuous | - |
| 5 | speed limit | SPL | continuous | mph |
| 6-10 | road function | RF | categorical nominal | - |
| 11 | crossing | C | categorical nominal (yes/no) | - |
| 12-16 | link end description | LED | categorical ordinal | - |
| 17 | roadside environment | RE | categorical nominal (res/non-res) | - |
| 18 | season | S | categorical nominal (summer/winter) | - |
| 19 | time of day | TOD | categorical nominal (morning/evening) | - |
| 20 | $PM_{10}$ background conc. | PM_BG | continuous | $\mu g/m^3$ |
| 21 | temperature | T | continuous | °C |
| 22 | temperature difference[a] | TD | continuous | K |
| 24 | wind speed | WS | continuous | m/s |
| 25 | maximal wind speed[a] | mWS | continuous | m/s |
| 26 | global radiation[b] | GR | continuous | $W/m^2$ |
| 27 | rain[b] | R | continuous | mm |
| 32-39 | wind direction | WD | categorical nominal | - |

**Table 5-1 Independent variables considered as potential determinants in multiple linear regression model**

a. disregarded due to near multicollinearity
b. disregarded due to insufficient data

The wind rose in Figure 5-1 shows all hourly average values for wind speed and wind direction that occurred during data collection. Since wind direction is a circular variable, measured in degrees from 0 to 360, the divergence of the values from zero does not increase continuously. The variable thus required a conversion to be useable as an independent variable in the multiple linear regression analysis. It was considered to convert the wind direction values into wind direction indices which would be values on a continuous scale. However, the approaches investigated did not lead to a satisfactory solution, as explained in Appendix C.2. It was therefore decided to treat wind direction (WD) as a categorical variable, coded with eight dummy columns, i.e. one category for each 45 degree section of the compass (cat0 (NNE): $0° < \alpha \leq 45°$, cat1 (ENE): $45° < \alpha \leq 90°$ etc.). The orange lines in Figure 5-1 give an indication of the proportion of measurements falling within the each of the dummy

categories defined for the variable. The maximum proportion of measurements (~ 28 %) lie in the fifth sector, indicated by the line equal to the radius in the figure.

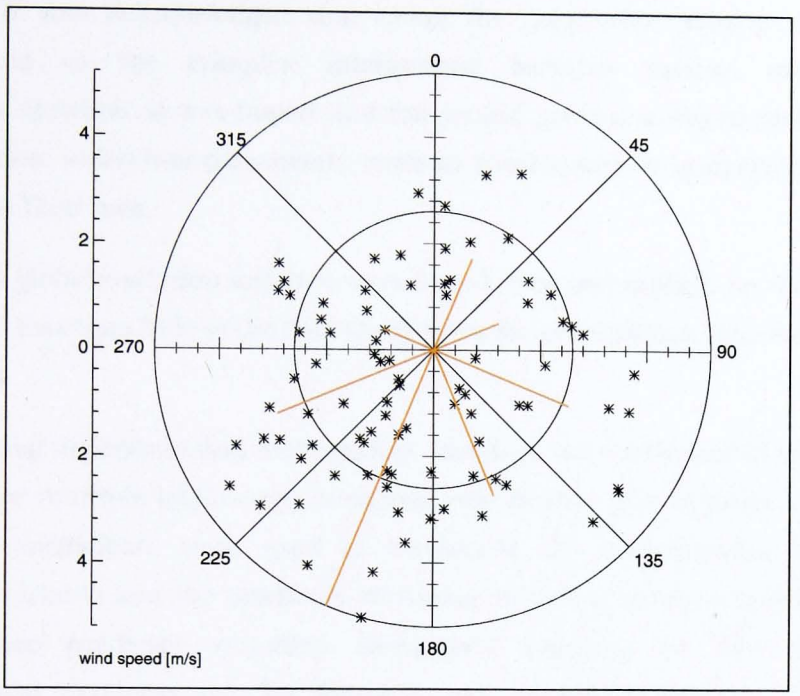


**Figure 5-1 Wind rose showing wind speed and direction in Leicester during all valid field study runs**

*The orange lines indicate the relative proportion of measurements falling within each of the eight dummy categories, i.e. maximum proportion = radius.*

This shows that, normal for the UK conditions, (south) westerly winds are observed most frequently while northerly winds are least often recorded. Findings from the analysis of the wind direction data, as described in Appendix C.2, additionally suggest that winds coming from the more often recorded directions tend to be stronger. To avoid extreme multicollinearity in the analysis, the first wind direction category was chosen as the reference category and excluded from the analysis, leaving seven dummy columns for this variable.

Traffic flow and speed data were also provided by Leicester City Council. They consisted of average values for morning and evening rush hour times for each link as output from the TRIPS[1] model, used by the City Council for traffic flow modelling across a subset of major roads for the city.

1. Transportation model (MVA, 1996).

Additionally, the two categorical nominal variables 'time of day' (TOD) and 'season' (S) were included to investigate whether diurnal or seasonal effects could be identified. Each variable was represented by one column, containing 0 (am, winter) or 1 (pm, summer). It was acknowledged that using the parameter 'season' results in a simplification of the complex interactions between various meteorological parameters. However, it was hoped that this would provide a way to identify general trends, in case individual parameters such as wind speed or temperature were not sufficient on their own.

The data on global radiation and rain were found to be incomplete, i.e. valid data was available for less than 50% of the field study records, and was therefore excluded from the analysis.

The remaining dependent and independent variables were analysed in terms of their suitability for multiple linear regression analysis. Scatter plot matrices and Pearson correlation coefficients were used to investigate the relationships between the dependent variable and the predictor variables as well as inter-correlation between the individual predictor variables. Scatterplot matrices for link specific and meteorological variables can be found in the appendix, Figures C-3 and C-4, respectively. Correlation tables are also included, see Tables C-1 and C-2.

Strong correlation (i.e. $r^2 > 0.5$) was observed between temperature (T) and temperature difference and between wind speed (WS) and maximal wind speed (mWS). These variables also showed rather high regression coefficients when regressed against the other variables (testing for near-extreme multicollinearity). Since the inclusion of strongly inter-correlated variables would make the multiple regression model unreliable, inter-correlated variables least correlated to the dependent variable (i.e. TD and mWS) were not included in the regression analysis. Moreover, URR showed a high multiple correlation coefficient when testing for multicollinearity and was strongly correlated to TRIPS traffic flow data (tFL). It was therefore also removed. The variables road function (RF) and link end description (LED) are categorical parameters. To avoid extreme multicollinearity the categories 'local' and 'open' were chosen as reference categories, respectively, and were excluded from the analysis. The two variables LED and RF were thus included in the analysis as four dummy variables each.

In order to ensure that a consistent data set was used in the analysis, all observations including extreme outliers for any of the independent variables or the dependent variable were removed from the data set. Most of the independent variables did not have such outliers, apart from: link length; TRIPS flow; temperature; and, wind speed. It was found that, by applying an appropriate upper and lower cut-off limit based on a percentile value, extreme outliers could be removed. For link length, TRIPS flow and wind speed, the 95th and 5th percentile were used as upper and lower cut-off values, respectively. For temperature it was sufficient to use the 98th and 2nd percentiles.

As mentioned above, normality plots for the dependent variable showed that the data more closely resembled a normal distribution after the removal of outliers, which is an important requirement for use of multiple regression analysis techniques. Although, this 'data cleaning' process may also lead to a considerably reduced data set, it was regarded as acceptable, considering the benefits of using more normally distributed data and minimising the potential bias due outliers. Moreover, the remaining data set still consisted of 1626 complete records, which is considered adequate for this type of analysis. However, it is important to note that the removal of the outliers reduced the variability of the analysed data set, which may impede the predictive ability of the regression models which are derived as part of the analysis. Scatterplots and correlation coefficients for the three dependent variables against the independent variables included in the models before backward elimination are included in Appendix C.3.

### 5.3.3 Multiple regression analysis

Based on the data set of 1626 observations, multiple linear regression with backward elimination was carried out for the three dependent variables ATE, ATEK and AC. The derived regression models are in the form of

$$Y = e^{b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_p X_p}. \qquad \textbf{(5-5)}$$

All models calculated during the backward elimination stages were tested for validity by inspecting probability and scatter plots to check whether the residuals were normally distributed and that the independent variables did not show any correlation with the residuals. The normality of the residuals was further assessed by performing

the Jarque-Bera test. Also, by plotting the unstandardized residuals against the predicted values for the dependent variable, it was verified that the modelled results fulfilled the assumption of equal variances (i.e. homoskedasticity). This was accompanied by calculating White's test for homoskedasticity. This test showed evidence of heteroskedasticity in all cases. Therefore, White's method of calculating robust standard errors was applied.

## 5.4 Results - Ultrafine particles

The analysis was first carried out for ungrouped data, i.e. the complete data set. Based on the main determinants identified for each dependent variable, the data set was the divided into sub-sets and the analysis run again for each sub-set.

### 5.4.1 Results for ungrouped data

The results of the multiple linear regression analyses for all three dependent variables are shown in Table 5-2. They include the multiple correlation coefficients for the initial models (i.e. with all variables included) and the parameters for the models after backward elimination. The most significant variables are marked in bold type face. All variables included after backward elimination were statistically significant ($p > 10^{-9}$). For more detailed tables showing the steps of the backward elimination process and the variables included at various significance levels, see tables in the appendix (Appendix C.4).

Main determinants - ATE and ATEK

The final models for all three dependent variables have multiple correlation coefficients $R^2 > 0.35$, which means that more than 35 % of the variability in the measured values can be explained by the models.

The same determinants were identified for ATE (average total exposure) and ATEK (average total exposure/km) with the first model showing a slightly higher multiple correlation coefficient of $R^2 = 0.39$. The main determinants for both models are, in order of importance: link end description (LED); link length (LL); and, meteorological conditions (i.e. season (S), wind direction (WD) and wind speed (WS)). Using $R^{2*}$ (ratio of between groups sum of squares to total sum of squares), the independent variables can be assessed in terms of their effect on the dependent variable. In the first model, the strong influence of LL on total exposure values ($R^{2*} = 0.11$) is to be

expected since it takes more time to complete longer links if all other parameters are the same, which leads to higher total exposure values.

| dependent variable | $R^2$ for compl. model | $R^2$ after BWE | variables included in final model | $b_0$ | $b_i$ | standard error | $R^{2*}$ |
|---|---|---|---|---|---|---|---|
| average total exposure (ATE) | 0.4412 | 0.3914 | link length (1) | 14.8157 | 1.9176 | 0.1668 | 0.1068 |
| | | | **link end description[a] (13) 'left turn'** | (0.0833) | **-0.0796** | **0.0676** | **0.2293** |
| | | | **(14) 'signal'** | | **0.5059** | **0.0358** | |
| | | | **(15) 'right signal'** | | **1.2361** | **0.0585** | |
| | | | **(16) 'right turn, no signal'** | | **1.2600** | **0.1508** | |
| | | | season (18) | | -0.3200 | 0.0370 | 0.0489 |
| | | | wind speed (24) | | -0.1691 | 0.0160 | 0.0412 |
| | | | wind direction[b] (33) 45°-90° (ENE) | | -0.0687 | 0.0871 | 0.0495 |
| | | | (34) 90°-135° (ESE) | | 0.0878 | 0.0731 | |
| | | | (35) 135°-180° (SSE) | | -0.3056 | 0.0724 | |
| | | | (36) 180°-225° (SSW) | | -0.2547 | 0.0619 | |
| | | | (37) 225°-270° (WSW) | | -0.3066 | 0.0620 | |
| | | | (38) 270°-315° (WNW) | | -0.3669 | 0.0672 | |
| | | | (39) 315°-360° (NNW) | | -0.2521 | 0.0927 | |
| average total exposure/km (ATEK) | 0.4162 | 0.3642 | link length (1) | 17.4383 | -2.6823 | 0.1629 | 0.0632 |
| | | | **link end description[a] (13) 'left turn'** | (0.0838) | **-0.0655** | **0.0675** | **0.1356** |
| | | | **(14) 'signal'** | | **0.5066** | **0.0354** | |
| | | | **(15) 'right signal'** | | **1.2528** | **0.0656** | |
| | | | **(16) 'right turn, no signal'** | | **1.2707** | **0.1507** | |
| | | | season (18) | | -0.3153 | 0.0367 | 0.0548 |
| | | | wind speed (24) | | -0.1674 | 0.0160 | 0.0504 |
| | | | wind direction[b] (33) 45°-90° (ENE) | | -0.0632 | 0.0893 | 0.0503 |
| | | | (34) 90°-135° (ESE) | | 0.0932 | 0.0726 | |
| | | | (35) 135°-180° (SSE) | | -0.3024 | 0.0720 | |
| | | | (36) 180°-225° (SSW) | | -0.2490 | 0.0621 | |
| | | | (37) 225°-270° (WSW) | | -0.2922 | 0.0624 | |
| | | | (38) 270°-315° (WNW) | | -0.3595 | 0.0676 | |
| | | | (39) 315°-360° (NNW) | | -0.2436 | 0.0931 | |
| average conc. (GM) (AC) | 0.4548 | 0.3973 | **temperature (21)** | 12.0875 | **-0.0244** | **0.0016** | **0.2486** |
| | | | wind speed (24) | (0.0363) | -0.1555 | 0.0102 | 0.1611 |
| | | | wind direction[b] (33) 45°-90° (ENE) | | -0.0910 | 0.0516 | 0.1489 |
| | | | (34) 90°-135° (ESE) | | 0.1125 | 0.0384 | |
| | | | (35) 135°-180° (SSE) | | -0.2500 | 0.0404 | |
| | | | (36) 180°-225° (SSW) | | -0.2725 | 0.0356 | |
| | | | (37) 225°-270° (WSW) | | -0.2621 | 0.0363 | |
| | | | (38) 270°-315° (WNW) | | -0.2542 | 0.0401 | |
| | | | (39) 315°-360° (NNW) | | -0.0894 | 0.0547 | |

**Table 5-2 Results of multiple linear regression with backward elimination**

a. reference category (13) 'left turn'

b. reference category (32) 0°-45° (NNE)

$R^{2*}$ ratio of the between groups sum of squares to the total sum of squares

The ATEK model has a lower correlation coefficient but still explains more than 35 % of the variability of the dependent variable. Interestingly, even though the actual length of the link has been cancelled out, link length still remains in the model as a significant predictor variable ($R^{2*}$ = 0.06). However, it is now negatively correlated to

the dependent variable, which means that high exposure values per km are expected for shorter links. The parameter with the strongest influence on both ATE and ATEK however is LED, showing the influence of junction layouts, in particular traffic lights, on observed exposure values. Of the included categories 'right signal' and 'right turn' have the strongest effect.

No published results from directly equivalent studies were available for comparison at the time of writing (2003 - 04). However, a personal exposure study by (Adams et al., 2001b), who used average $PM_{2.5}$ journey exposure levels, derived a model that explained about 34 % of the variability in exposure values, with 'Route' being the most influential factor, explaining more than 20 %. This could be representing the influence of queuing similarly to the variable 'link end description' used in the current study, with $R^{2*} = 0.23$ in the ATE and $R^{2*} = 0.14$ in the ATEK model. The finding that link length is negatively correlated to total exposure/km supports the hypothesis that queuing is a significant factor for exposure levels. Since road links are by definition typically terminated by a junction (i.e. change of flow or number of lanes), short links are associated with proportionally more queuing time than long links. The variables relating to meteorological conditions included in the two models were season (S), wind direction (WD) and wind speed (WS). Even though they are less influential than LED and LL, this confirms that meteorological conditions affect exposure values (see discussion below).

Main determinants - AC

The main determinants identified for AC (average concentration) do not include any road specific parameters but are solely based on meteorological conditions, i.e. temperature, wind speed and wind direction. Contrary to the results for $PM_{2.5}$ presented by Adams et al. (2001b) however, wind speed is less influential than temperature, $R^{2*} = 0.16$ and $R^{2*} = 0.25$ respectively, which supports the hypothesis that exposure to ultrafine particles does not follow the same overall pattern as exposure to particles in the $PM_{2.5}$ range.

The variables wind speed and wind direction are included in all three models. The negative coefficient for wind speed suggests that dispersion of pollutants lowers driver exposure by diluting emissions before they enter the vehicle, despite the close proximity to the emission sources, i.e. vehicle exhausts. The observed association between lower wind speeds and elevated concentration and exposure values is in

agreement with results from studies by (Harrison et al., 1997) and (Deacon et al., 1997) who identified strong negative correlations between outdoor concentrations of $PM_{2.5}$ and wind speed.

The wind direction categories in the three models all have negative regression coefficients apart from cat0 (0-45°), which is the reference category and therefore has a coefficient of 0, and cat2 (90°-135°), with coefficients ranging from 0.07 (ATE) to 0.11 (AC). For all models cat1 has a comparatively small negative coefficient as has cat8 for AC. This means that elevated concentration and exposure values are observed on days with prevailing winds coming from easterly (and for AC northerly) directions, and particularly ESE. It is not immediately clear why this should be. The wind rose in Figure 5-1 shows that this category does not contain a particularly low or high number of values. It is also unlikely that it is due to the influence of a strong particle source in ESE direction from the study area since ultrafine particles concentrations are mainly dependent on local emissions, e.g. other vehicles in the traffic stream (Harrison et al., 1999a). However, as mentioned in Section 5.3, the weak negative correlation found between wind speed and the 'normalized' wind direction indicates that winds coming from the most frequently recorded wind directions tend to be stronger. The most frequently occurring wind direction for the field study was 199°, approx. SSW, i.e. about opposite the WD categories with the weak negative or positive relationships to the dependent variable. This indicates that the winds from easterly directions are likely to be weaker and thus confirms that low wind speeds are associated with higher concentration and exposure values.

Temperature in the AC model has a negative regression coefficient, i.e. low temperatures are linked to high concentration values. Although, as far as is known, no such relationships between exposure or concentrations and temperature has been reported in personal exposure studies. However, as discussed in Section 4.2, many studies which investigated particle concentrations (not in-vehicle), found that elevated concentrations were typically observed at cooler temperatures, i.e. morning hours and winter months. This is thought to be partly due to ambient temperatures affecting particle transformation and/or cold start emissions, leading to elevated concentrations at lower temperatures (Kittelson et al., 2000; AEAT, 2001). It is also possible that the parameter temperature was identified as important because it indirectly represents the effect of other seasonal and diurnal conditions on

concentration values, such as reduced dispersion during the mornings and winter months, and stronger morning than evening rush hour traffic flows, leading to higher concentrations during periods with comparatively low temperatures (Molnár et al., 2002; Harrison et al., 1999). If the parameter temperature in the model does indeed represent the effects of seasonal and diurnal influences, the parameters season and time of day would be expected to have been identified as main determinant also. That this is not the case may be due to the relationships between temperature and these parameters being rather weak and/or the choice of variable categories (i.e. two season categories may not be sufficient).

However, the results from the ATE and ATEK models show that season is indeed a main determinant of exposure (if not average concentration). Season, which takes the values 0 for winter and 1 for summer, is in both cases negatively correlated to the dependent variable, thus associating elevated exposure values to winter conditions.

The fact that the final average concentration model (AC) only includes temperature, wind speed and wind direction emphasises the difference between 'concentration' and 'exposure'. Whereas average concentrations are mainly affected by meteorological conditions, exposure values seem to be more strongly related to road specific parameters. However, if concentration values were to be used for exposure assessment, factors such as 'queuing time', which has been identified as a determining factor by the other two models, would have to be considered, i.e. the calculation of exposure values from concentration averages would have to take the time spent on a link or route into account. This time factor and hence the exposure result would be higher for links with longer queuing periods, with all other parameters being the same.

Variability explained by the models

All three models have multiple correlation coefficients greater than 0.35, which means that they explain more than 35 % of the variation in the dependent variables. This is a reasonably good fit considering that the time series plots shown in Section 4.3 indicate that in-vehicle UF concentrations are strongly influenced by the emissions from the vehicle in front and that no parameter representing this was included in the analysis. As a study by Johnson and Ferreira (2001) shows that the inclusion of such a parameter may well have led to a better model. Their study aimed to predict levels of sub-micrometer particles contributed to road side concentrations by traffic flows.

They found that their basic model's accuracy could be increased from 58 % to 75 % by incorporating fleet composition information in addition to temperature and wind speed measurements. Although, details regarding the vehicle directly ahead of the instrumented vehicle were recorded during field study measurements, it was not practical to include the data in this analysis due to the time required to manually extract this information from the logs for a large number of runs. Moreover, such data is highly run specific and is thus not suitable for the use in this multiple regression analysis, which aimed to identify determinants of driver exposure which could be used to produce a generalizable model, as explained in Section 6.1.1. Similarly, characteristics of traffic flows and fleet composition on individual road links can only be derived from extensive traffic counts, and do not therefore fulfil the criterion of being 'readily available', which was a requirement for all variables used in this analysis and modelling exercise (see also Section 6.1).

### 5.4.2 Results for grouped data

In order to investigate whether the models could be improved by using separate model equations for grouped data, e.g. using summer or winter data only to calculate coefficients, sub-sets of data were subjected to further multiple regression analyses. It was thought that this may also reveal effects of other determinants which may have been masked when analysing the complete data set. The sub-sets were compiled by grouping observations from the complete data set according to the main determinants identified in the first step, i.e. LED and meteorological conditions.

Methods

For grouping data depending on the LED, the five dummy columns of this variable where condensed into one, with the categories 'open' and 'left turn' coded as 0 (LED_0) and the categories 'signal', 'right signal' and 'right turn' coded as 1 (LED_1). This division was based on the results from the first regression analysis that found a negative regression coefficient for 'left turn' and positive coefficients for the other three categories. Since the category 'open' had been left out as a reference category, its coefficient can be seen as being zero and it was therefore grouped together with 'left turn'.

The variables related to meteorological conditions identified in the first analysis included different combinations of season, wind speed, wind direction and

temperature. This suggests a general influence of meteorological conditions on exposure values, without one of the variables being distinctly more influential than the others. Since it was not clear which limit value to choose in order to be able to group data by temperature, wind speed or wind direction, it was decided to group them by season. This decision was based on the assumption that seasonal variations in meteorological conditions affect the three other variables. Grouping data by season therefore provided a suitable way of investigating the effect of the meteorological conditions on exposure and concentration values by using just one further sub-division of the data.

It was also considered to further divide the data sets into 'summer/LED_1', 'summer/ LED_0' etc. in order to investigate parts of the data set further, but the number of observations in these groups was too small to ensure statistically meaningful results. Moreover, using data sets with small numbers of observations would also have led to the exclusion of one or more individual categories for the categorical variables due to singularity (i.e. all observations in the data set having the same value for that category), which may have biased the results in favour of the non-categorical variables.

Separate regression analyses were carried out for the various sub data sets in a similar manner as described above, i.e. multiple regression with backward elimination. The multiple correlation coefficients as well as the variables identified as most significant for each sub-set are shown in Table 5-3, with the most influential marked in bold. The results for the models based on the complete data set are also shown for comparison. All variables included in the models after backward elimination were statistically significant ($p < 10^{-9}$).

| | | ATE | | ATEK | | AC | |
|---|---|---|---|---|---|---|---|
| **sub data set** | **N** | **$R^2$** | **variables** | **$R^2$** | **vars** | **$R^2$** | **vars** |
| **all** | 1626 | 0.3914 | LL, **LED**, S, WS, WD | 0.3642 | LL, **LED**, S, WS, WD | 0.3973 | T, **WS**, WD |
| **LED_0** | 627 | 0.3399 | LL, **RF**, C, WS, WD | 0.4526 | LL, **RF**, S, WS, WD | 0.3714 | **T**, WS, WD |
| **LED_1** | 999 | 0.2937 | **LL**, RF, C, T, WS | 0.2964 | LL, RF, C, **T**, WS | 0.4579 | RF, **T**, WS, WD |

**Table 5-3 Results for the multiple regression analyses using sub-sets of data, showing multiple correlation coefficients and main predictor variables for the final models after backward elimination**

| | | ATE | | ATEK | | AC | |
|---|---|---|---|---|---|---|---|
| **summer** | 1101 | 0.3676 | LL, **LED**, PM_BG, WS | 0.3319 | LL, RF, **LED**, WS | 0.2993 | PM_BG, **WS**, WD |
| **winter** | 525 | 0.3822 | LL, **LED**, WS, WD | 0.3781 | LL, **LED**, WS, WD | 0.5051 | **T**, WS, WD |

**Table 5-3 Results for the multiple regression analyses using sub-sets of data, showing multiple correlation coefficients and main predictor variables for the final models after backward elimination**

Results and discussion

The results show that models with higher multiple correlation coefficients can be found using sub-sets of data. For observations grouped by LED, the main additional determinants identified for links in category LED_0 include road function (RF) and crossing (C) whereas the models for links in category LED_1 include temperature (T) rather than wind direction (WD), which had been identified as a main determinant for the complete data set. This shows that physical link parameters in general have an effect on exposure values but, due to the complex underlying relationships, it is difficult to attribute changes in exposure to one particular parameter.

The models derived for data grouped by season generally confirmed the main determinants identified earlier, with RF and PM10_BG (for ATEK_summer and AC_summer, respectively) being the only exceptions. The correlation coefficients for the winter models were similar or higher than for the all inclusive models. The most pronounced change was observed for AC, where the winter model explains approximately 27 % more of the variation in exposure data than the all inclusive model.

Although some models with higher multiple correlation coefficients were found using sub-sets of data, the results show that improvements are usually only observed for one of the sub-sets whereas the model for the second one has a lower multiple correlation than the general model which is based on the whole data set. This suggests that this general model represents a good compromise, providing a reasonably well correlated model which can be applied to all data.

## 5.5 Results - $PM_{10}$ and $PM_{2.5}$

For comparison, and in order to investigate to what extent driver exposure to larger (i.e. more commonly measured) particle size fractions can be used as a surrogate for ultrafine particle exposure, multiple linear regression analyses were also carried out

using measured $PM_{10}$ and $PM_{2.5}$ mass concentration data as the dependent variables. The same methods were used for data preparation, variable selection and regression calculations as described above for ultrafine particles.

The results of these analyses are presented in Table 5-4. Results for ultrafine particles have been included for comparison. The number of observations was 1704 for the $PM_{10}$ and $PM_{2.5}$ and 1626 for UF. All variables included in the final models were statistically significant ($p < 10^{-9}$). The most significant variable for each model is denoted in bold type face.

The observed multiple correlation coefficients for the $PM_{2.5}$ and $PM_{10}$ models are similar to the ones reported for UF, which, however, tend to be slightly higher. Ranging from 0.33 to 0.44 they are comparable to the results obtained by (Adams et al., 2001b) who reports that up to 34 % of the variation in exposure values ($PM_{2.5}$) could be explained by a model including variables for wind speed, travel mode and route factors.

| | ATE | | | ATEK | | | AC | | |
|---|---|---|---|---|---|---|---|---|---|
| **size range** | $R^2$ | Predictor variables included | $R^{2*}$ | $R^2$ | Predictor variables included | $R^{2*}$ | $R^2$ | Predictor variables included | $R^{2*}$ |
| **$PM_{10}$** | 0.3468 | LL<br>tSP<br>**LED**<br>PM_BG | 0.1123<br>0.0087<br>**0.2244**<br>0.0497 | 0.3301 | LL<br>**RF**<br>LED<br>PM_BG | 0.2375<br>**0.2984**<br>0.1616<br>0.2900 | 0.3429 | **PM_BG**<br>T<br>WS<br>WD | **0.1517**<br>0.0010<br>0.1288<br>0.1212 |
| **$PM_{2.5}$** | 0.3738 | LL<br>**LED**<br>PM_BG<br>WD | 0.0880<br>**0.1905**<br>0.1145<br>0.0790 | 0.3632 | LL<br>LED<br>PM_BG<br>**WD** | 0.2661<br>0.2231<br>0.2909<br>**0.3204** | 0.4407 | **PM_BG**<br>WS<br>WD | **0.2644**<br>0.1150<br>0.2060 |
| **UF** | 0.3914 | LL<br>**LED**<br>S<br>WS<br>WD | 0.1068<br>**0.2293**<br>0.0489<br>0.0412<br>0.0495 | 0.3642 | LL<br>**LED**<br>S<br>WS<br>WD | 0.0632<br>**0.1356**<br>0.0548<br>0.0504<br>0.0503 | 0.3973 | **T**<br>WS<br>WD | **0.2486**<br>0.1611<br>0.1489 |

**Table 5-4 Multiple linear regression results for three particle size ranges, showing multiple correlation coefficient $R^2$ and $R^{2*}$ (ratio of the between groups sum of squares to the total sum of squares) for variables included in 'best models'**

The results show that the main determinants for driver exposure to particles of the larger size fractions are similar to the ones identified for ultrafine particle exposure. For all three size fractions, link length (LL) and link end description (LED) are included in the final ATE and ATEK models. The significant difference, however, is that the third variable included in the two models for the larger size fractions is $PM_{10}$ background concentration (PM_BG), while the UF models use season (S) and wind

speed (WS) instead. This means that $PM_{10}$ background concentrations may be used as a basis to predict driver exposure to larger particles, but they do not have a statistically significant relationship with driver exposure to smaller particles. This is in agreement with results from a large number of studies (e.g. Weijers et al., 2001; Harrison et al., 1999) which found that mass concentration measurements are not suitable for the evaluation of personal exposure to particulate matter from transport sources.

Interestingly, the dependence of exposure values on meteorological parameters, which has also been reported by Adams et al. (2001b), was confirmed by wind speed being included in the ATE and ATEK models for $PM_{2.5}$. However, the equivalent $PM_{10}$ models did not include any meteorological parameters, whereas the UF model includes wind speed as well as season and wind direction. Particle concentrations in the $PM_{2.5}$ range can therefore be said to have characteristics of both, the larger size ranges (i.e. dependence on background concentrations) as well as the smaller particles (i.e. susceptibility to wind direction). They may therefore be slightly more suitable than $PM_{10}$ concentrations as indicators for ultrafine particles. Moreover, the strong significance of the road layout parameters in the $PM_{10}$ and $PM_{2.5}$ models, which affect queuing time, may have masked the less significant effect of wind speed on exposure values. This is supported by the fact that in the average concentration models (AC), where the influence of queuing time is practically removed, wind speed is one of the more influential parameters. In fact, wind speed is included as a significant variable with a negative correlation in all average concentration models, which is in agreement with studies by (Harrison et al., 1997; Deacon et al., 1997) who identified strong negative correlations between outdoor concentrations $PM_{2.5}$ and wind speed. As discussed above, the inclusion of wind direction as a main determinant can be linked to its relationship with wind speed and thus further confirms that low wind speeds contribute to elevated particle concentrations.

In conclusion, it can be said that the main determinants for driver exposure to $PM_{10}$ and $PM_{2.5}$ particles include the same link specific parameters as the UF model, but are also able to make use of $PM_{10}$ background concentration measurements. However, these background measurements are not useful for the prediction of driver exposure to ultrafine particles, which seems more closely related to meteorological parameters. The results for the AC models confirm that meteorological parameters

have a strong effect on average concentrations for all particles. Particularly temperature/season seems to influence ultrafine particle concentrations and exposure, with higher values expected at lower temperatures.

## 5.6 Summary

The results from the multiple regression analysis confirmed the observations made in Chapter 4 that meteorological and road parameters affect the concentrations of ultrafine, $PM_{10}$ and $PM_{2.5}$ particles. The analysis identified 'link length' and 'link end description' (i.e. type of traffic light set-up terminating a link) as the main determinants of driver exposure to ultrafine particles. In addition, the parameters 'season', 'wind speed' and 'wind direction' were identified as significant predictor variables, confirming the influence of meteorological parameters on driver exposure. Together, these parameters explained more than 36 % of the variability in link exposure values. Average link concentrations of ultrafine particles, however, were shown to be mainly influenced by meteorological parameters, with 'temperature', 'wind speed' and 'wind direction' alone leading to a model that explained close to 40 % of the variability in observed average link concentration values.

The multiple regression results for $PM_{10}$ and $PM_{2.5}$ suggest that exposure to particles of these size ranges is mainly influenced by road specific parameters, i.e. 'link length', 'link end description', 'road function' (only $PM_{10}$) and '$PM_{10}$ background concentrations'. For $PM_{2.5}$ exposure, 'wind direction' was also found to be a significant factor. The main predictor variables for average link concentrations for both particle size ranges were found be meteorological parameters, i.e. similar to results for UF, and '$PM_{10}$ background concentrations'.

Comparison of the results from the regression analysis for the different size ranges indicates that they are affected differently by external parameters. Although link exposure for all size ranges is affected by road specific parameters, link exposure to ultrafine particles is also strongly dependent on meteorological parameters, while $PM_{10}$ exposure can be linked to $PM_{10}$ background concentrations. $PM_{2.5}$ exposure takes up middle ground, as is illustrated in Figure 5-2. This further supports the hypothesis that the mass concentration data cannot be used to infer ultrafine particle exposure levels.

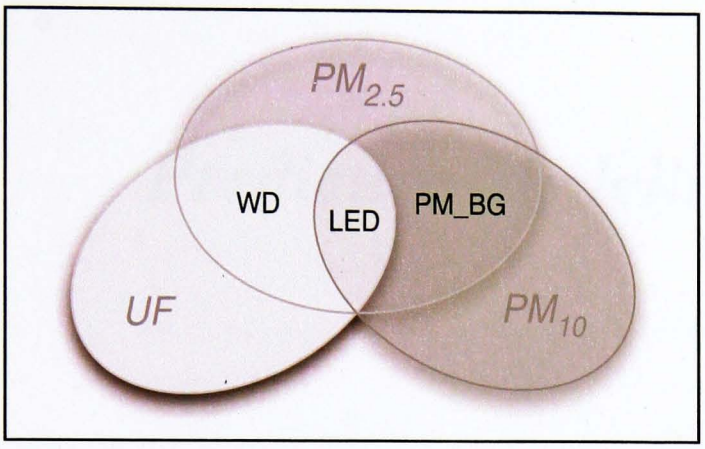


**Figure 5-2 Venn diagram illustrating the similarities between results for ultrafine particles, $PM_{2.5}$ and $PM_{10}$**

It is important to note, however, that the results presented here may have limited potential for extrapolation due to the design of the field study design. In particular the assumption that particle concentrations measured while driving along a road link are influenced by traffic conditions on that link, rather than the previous link(s), may have affected the detection of road specific influences on particle concentrations. Moreover, the focus of the data collection on a limited range of road types (urban) and traffic conditions (rush hour) is likely to have biased the results to some extent, i.e. increasing the likelihood of meteorological parameters to be identified as the main determinants by limiting the range and thus variability of road and traffic related conditions.

During the first part of the data analysis (Chapter 4) various aspects of driver exposure were explored, using the field study data to gain an understanding about the short-term variability of in-vehicle concentrations and the possible underlying relationships with external parameters. In the second part of the analysis, i.e. this chapter, statistical methods were used to formally identify the main determinants of driver exposure. The first part of the next chapter aims to investigate the potential of some of these findings in terms of exposure estimation by utilizing the field study data to predict in-vehicle link and journey exposure levels.

Chapter

# 6 Predictive Modelling

The previous chapters have explored the field study data in order to gain an understanding of the dynamics and main determinants of in-vehicle concentrations of particulate matter and thus driver exposure. The analysis in Chapter 4 showed how average exposure and total exposure levels varied for different field study journeys (Sections 4.2 and 4.7 respectively). However, it is desirable to utilize the findings from this study to estimate the exposure of commuters on other journeys. Such estimates could then be used to assess the contribution of exposure from car commuter journeys to overall exposure levels. Alternatively, it may be possible to derive an annual average value for the exposure in urban transport environments, which could be used to estimate the exposure of commuters based on the time travelled rather than on the individual journeys.

This chapter explores the use of the field study for driver exposure modelling. Two different approaches are investigated. First, the models derived from the multiple linear regression analysis (Chapter 5) are used to predict link and journey exposure. Limitations and possible areas of applications are explored. The second part of the chapter investigates the field data's potential for annual exposure modelling. A method is proposed which utilizes field study measurements together with meteorological reference data to estimate average annual exposure.

## 6.1 Journey exposure modelling

As outlined in Section 2.3, the assessment of personal exposure requires knowledge about pollutant concentrations in the microenvironments of interest as well as time-

activity pattern information. Personal exposure studies, therefore, tend to be based on measurements of pollution data from personal monitors which are accompanied by time-activity logs for the measurement period. As a result, these measured values only provide direct information about personal exposure during the measurement period. However, by applying appropriate statistical methods, these data can, in principle, be used to generalize about personal exposure in similar situations and over long time periods.

### 6.1.1 Objectives and models

The aim of the modelling exercise was to investigate whether data such as those collected during the field study can be used to develop a model that predicts driver exposure for commuter journeys on an urban road network based on readily available parameters. Only parameters that can generally be obtained without additional field studies were considered, such as meteorological data, $PM_{10}$ background concentrations, modelled traffic flow and speed data, in order ensure that the models can be easily applied without requiring time consuming data collection.

As outlined in the previous chapter, multiple linear regression analysis (MLR) is often used as a tool to investigate the complex underlying influences of parameters affecting pollutant concentrations and to identify the ones with the strongest effects (Adams et al., 2001b). The regression models generated by such an analysis can also be used to predict values for the dependent variable (model output) from a number of independent variables (model input) (e.g. Chaloulakou et al., 2003; Johnson and Ferreira, 2001). This approach is used here to estimate the exposure to particulate matter on individual road links and whole journeys.

In order for results from the multiple linear regression analysis to be used for the modelling of link or journey exposure, it is important to first identify which parameters would be best suited to represent driver exposure to particulate matter in an analytical model, i.e. as the dependent variable. Three scenarios have been identified that allow the calculation of driver exposure for road links and journeys, using multiple linear regression models and readily available input data.

A schematic overview of the three approaches is given in Figure 6-1. The phases **A** to **D** of the modelling process are indicated at the bottom of the figure. All three scenarios are based on the idea of predicting driver exposure for a journey by

modelling exposure on individual road links and then combining these results to generate an estimate for total exposure on that journey. For this purpose, the journey route is split up into road links with constant physical and dynamic parameters (Phase **A**) as described in Section 3.1.

In order to predict total exposure on the individual links, the relevant input data and model equation are then used to generate the intermediate predicted values, i.e. output from the intermediate model (Phase **B**). By applying a conversion calculation, i.e. multiplication with a further parameter, these values are then converted to provide a predicted total exposure value for each link (Phase **C**). In a final step, the exposure values for the individual links are added up to provide exposure estimates for the whole journey (Phase **D**).

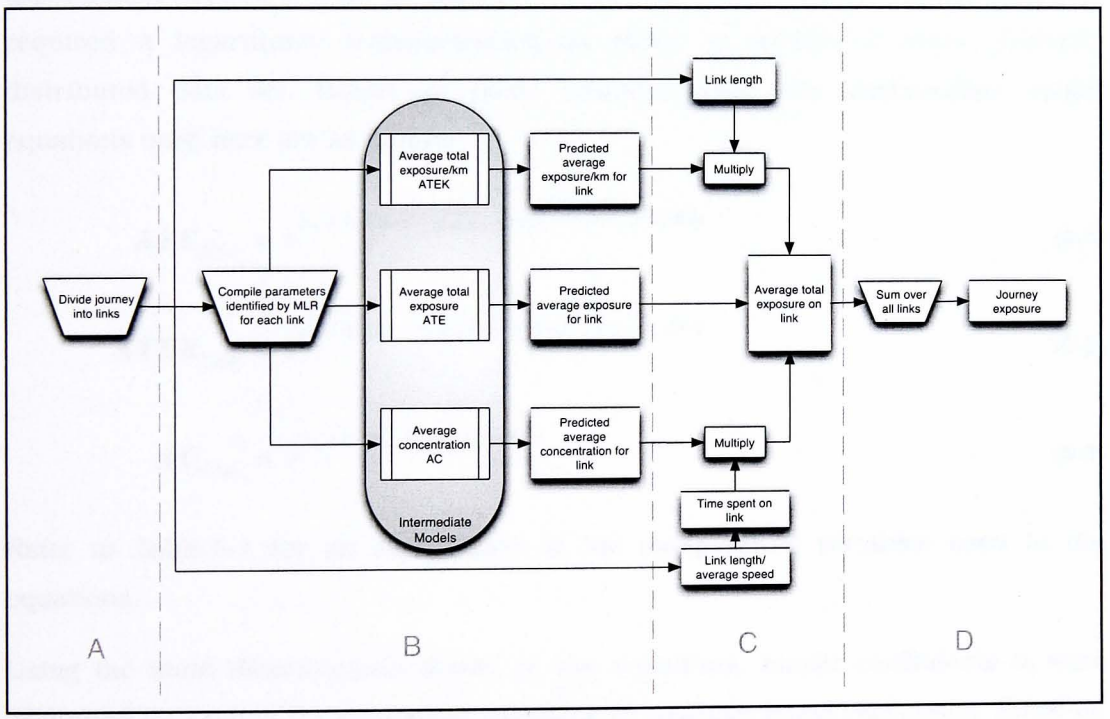


**Figure 6-1 Scenarios for link and journey exposure modelling**

The parameters chosen as output variables from the intermediate models are average concentration (AC), average total exposure (ATE) and average total exposure/km (ATEK). The AC and ATE models were chosen in order to investigate whether more accurate predictions can be achieved by modelling average concentrations for a link and multiplying them by the time spent on that link or by modelling total exposure values

directly. ATEK was identified as an alternative modelling approach since it provides exposure values per km, which can then be multiplied by link length in order to generate total link exposure values.

### 6.1.2 Methods

In terms of regression modelling, each of the output parameters from the intermediate models (Phase **B**) is equivalent to the dependent variable (Figure 6-1). As outlined in Section 5.3 (Table 5-1), the independent variables initially considered for the model development included road specific parameters (e.g. link length, road side environment etc.) and meteorological parameters (e.g. temperature, wind speed etc.). The analysis in Chapter 5 identified which of these variables are the most significant, i.e. have the strongest influence on the associated dependent variable (Table 5-2). As explained in Section 5.3, the strong skew of the dependent variables meant that they required a logarithmic transformation in order to achieve a more normally distributed data set. Based on these considerations, the intermediate model equations used here are as follows.

$$ATE_{link} = e^{b_0 + b_1 LL + b_2 LED + b_3 S + b_4 WS + b_5 WD} \quad \text{(6-1)}$$

$$ATEK_{link} = e^{b_0 + b_1 LL + b_2 LED + b_3 S + b_4 WS + b_5 WD} \quad \text{(6-2)}$$

$$AC_{link} = e^{b_0 + b_1 T + b_2 WS + b_3 WD}. \quad \text{(6-3)}$$

Refer to Table 5-1 for an explanation of the independent variables used in the equations.

Using the main determinants shown in the equations, model coefficients $b_i$ were generated for each of the dependent variables by multiple linear regression. Since no test data set was available for model evaluation purposes, the data set was split into two sets by randomly selecting 10 % of the data and excluding it from the multiple regression analysis. The multiple regression calculations thus utilized 90 % of the data to derive the model coefficients. The model equations were then applied to the remaining 10 % of the data to calculate the predicted values for the dependent

variable. A comparison between these predicted values and the associated observed values then formed the basis of the model evaluation process.

However, using this method, misleading evaluation results may have been produced in cases where the randomly selected evaluation data happened to consist of values which did not truly represent the variability and spread inherent in the whole data set. In order to reduce the chance of such bias skewing the results, these calculations were repeated ten times, each time randomly selecting the 90 % of observations to be included anew. In each case the calculated model equations were applied to the remaining 10 % of the data and a comparison between observed and predicted values was carried out. Descriptive statistics of the evaluation parameters calculated for these ten cases were then compiled to aid the assessment of the variation in the models' results.

### 6.1.3 Evaluation of model performance - Theory

The parameters used for the model evaluation were chosen as follows. The predictive ability of a model can be evaluated visually by inspecting probability and equal variance plots of the residuals, as described in Section 5.2. In addition to the visual assessment, various statistical moments can be calculated. For example, a basic approach was used by Ebelt et al. (2000) who calculated the residual sum of squares (*RSS*) divided by the number of samples to assess the difference in variability between observed and predicted values. A more thorough analysis has been employed by Chaloulakou et al. (2003) who also reported (in addition to basic descriptive moments such as the means and standard deviations of observed and predicted values) the mean square error (*MSE*), the mean bias error (*MBE*), the root mean square error (*RMSE*) as well as the intercept $a$ and the slope $b$ in

$$\hat{P} = a + bO_i \tag{6-4}$$

where $O$ stands for the observed values and $\hat{P}$ for the expected predicted values, based on the least-squares regression between observed and predicted values. These moments have been used in many other studies in conjunction with the conventional Pearson's product-moment correlation ($r$) or its square, the coefficient of determination ($r^2$), in order to compare the model performances. Willmott (1982) provides a comprehensive summary of these indices, their importance, relationships

and interpretation. He recommends to use for model assessment purposes the averages of observed and predicted values ($\overline{P}$ and $\overline{O}$), their standard deviations ($s_p$, $s_o$) together with *MAE* (mean absolute error) and *RMSE*, which take the form of

$$MAE = \frac{\sum_{i=1}^{N} |P_i - O_i|}{N} \tag{6-5}$$

and

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (P_i - O_i)^2}{N}}, \tag{6-6}$$

where $N$ is the number of observations. $\overline{P}$, $s_p$, $\overline{O}$ and $s_o$ generally describe the two variates and *MAE* is a descriptive index for the average difference between observed and predicted values. *RMSE* also provides an indication of this difference but is more sensitive to extreme values. The difference between *MAE* and *RMSE* results from the weighting of each $(P_i - O_i)$ by its square which can inflate the *RMSE*. The root mean square error can be regarded as a high estimate of *MAE*. In order to assess how well the model explains the major trends and patterns present in observed data, the *RMSE* can be disaggregated into a systematic and an unsystematic portion as follows

$$RMSE_s = \sqrt{\frac{\sum_{i=1}^{N} (\hat{P}_i - O_i)}{N}} \tag{6-7}$$

$$RMSE_u = \sqrt{\frac{\sum_{i=1}^{N} (P_i - \hat{P}_i)}{N}}. \tag{6-8}$$

A model can be considered 'good' if the systematic difference approaches zero and the unsystematic difference approaches the *RMSE*. Since the system is conservative for the squared values of these measures (i.e. $MSE = MSE_s + MSE_u$) they can be used to calculate the proportions of *MSE* arising from systematic and unsystematic errors contained in the model.

In order to assess whether a model generally over- or underpredicts, the mean bias error (*MBE*) can be used, which is defined as follows

$$MBE = \frac{\sum_{i=1}^{N} (P_i - O_i)}{N}. \qquad \textbf{(6-9)}$$

Willmott (1982) also points out that $r$ or $r^2$ are less informative than the slope $b$ and the intercept $a$ of the least-squares regression for the assessment of the nature of the linear covariance between $P$ and $O$. He further introduces the 'index of agreement' ($d$), which is particularly suitable for making cross-comparisons between models since it provides an indication about the relative size and the nature of the average differences, and is described by

$$d = 1 - \frac{\sum_{i=1}^{N} (P_i - O_i)^2}{\sum_{i=1}^{N} (|P'_i| - |O'_i|)^2}, \quad 0 \le d \le 1 \qquad \textbf{(6-10)}$$

where $P'_i = P_i - \overline{O}$ and $O'_i = O_i - \overline{O}$. This index $d$ has been used in various studies in order to compensate for the inadequacy of $r^2$, which is often not in agreement with the patterns and trends reported by the included difference measures (e.g. Kukkonen et al., 2001; Chaloulakou et al., 2003). It is interesting to note that according to Kukkonen et al., (2001) "numerical experiments show that a totally random predicted concentration time series, having the same range of variability as the measured time series, would result in an IA [Index of Agreement] value equal to approximately 0.40".

The model evaluation was carried out in three stages. First, the predictive ability of the intermediate models was investigated, i.e. predicting values for the dependent variables ATE, ATEK and AC (results from Phase **B**). Secondly, these output values of the intermediate models were converted to link specific average total exposure values as shown in Figure 6-1 so that the modelled results for all scenarios could be compared directly (output from Phase **C**). Finally, an assessment of the models in terms of journey exposure prediction (results of Phase **D**) was carried out. The findings from the three model evaluation stages are outlined and discussed in the following sections.

### 6.1.4 Intermediate models

Using the described methods, model evaluation statistics were calculated for the three intermediate models for ATE, ATEK and AC (Eq 6-1 to Eq 6-3). In order to investigate whether more specific models would provide more accurate models, the best models based on grouped data (i.e. the models with the strongest predictive ability) were also investigated. As explained in Section 5.4.2, the best grouped data models for ATE and AC were based on winter data only, (i.e. ATE(W) and AC(W), respectively). The grouped model for ATEK, ATEK(LED_O), only used data collected on links which are either terminated by a left turn or end at a junction without traffic signals (i.e. where link end description was coded as 0 in the multiple regression analysis). Based on the main determinants for those models, which are listed in Table 5-3, three additional modelling equations were defined as follows

$$ATE(W)_{link} = e^{b_0 + b_1 LL + b_2 LED + b_4 WS + b_5 WD} \quad \text{(6-11)}$$

$$ATEK(LED\_0)_{link} = e^{b_0 + b_1 LL + b_2 RF + b_3 S + b_4 WS + b_5 WD} \quad \text{(6-12)}$$

$$AC(W)_{link} = e^{b_0 + b_1 T + b_2 WS + b_3 WD}. \quad \text{(6-13)}$$

For more details on the dependent and independent variables and the statistical methods used to derive these models refer to Chapter 5.

Summary statistics for the ten model calculation results for all investigated data sets are shown in Table 6-1. The first two rows refer to the model equations derived from the multiple regression calculations, followed by descriptive statistics for observed and predicted values, and results of the performance tests for each model as proposed by Willmott (1982) and explained above. For each statistical index, the mean is given for the results from the ten calculation runs, together with its standard deviation. Due to the way they are calculated, the model comparison parameters in the lower part of the table can only be used to compare models for the same dependent variable type (e.g. ATE to ATE(W), ATEK to ATEK(LED_O), etc.), apart from $r^2$ and the index of agreement, $d$, which has been designed to allow cross-comparisons between models.

| Model | ATE[a] (N=10) | | ATE(W)[a] (N=10) | | ATEK[b] (N=10) | | ATEK(LED_0)[b] (N=10) | | AC[c] (N=10) | | AC(W)[c] (N=10) | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | mean | stddev | mean | stddev | mean | stddev | mean | stddev | mean | stddev | mean | stddev |
| $R^2$ | 0.390 | 0.003 | 0.381 | 0.018 | 0.360 | 0.004 | 0.457 | 0.010 | 0.396 | 0.003 | 0.506 | 0.011 |
| RSS | 622.87 | 8.20 | 173.33 | 5.64 | 612.12 | 8.08 | 180.38 | 4.53 | 210.23 | 1.78 | 42.59 | 1.13 |
| obs. values | N=163 | | N=53 | | N=163 | | N=63 | | N=163 | | N=53 | |
| $GM_o$ | 2.58E+06 | 9.23E+04 | 3.70E+06 | 3.70E+05 | 1.23E+07 | 4.38E+05 | 8.84E+06 | 8.01E+05 | 7.20E+04 | 2.84E+03 | 9.20E+04 | 5.19E+03 |
| $stddev_o$ | 2.27 | 0.13 | 2.24 | 0.11 | 2.24 | 0.08 | 2.07 | 0.10 | 1.64 | 0.04 | 1.58 | 0.07 |
| $median_o$ | 2.56E+06 | 1.26E+05 | 3.82E+06 | 5.41E+05 | 1.19E+07 | 9.67E+05 | 8.42E+06 | 8.95E+05 | 7.16E+04 | 3.21E+03 | 9.99E+04 | 7.39E+03 |
| $min_o$ | 3.04E+05 | 1.22E+05 | 7.04E+05 | 1.49E+05 | 2.01E+06 | 5.00E+05 | 2.10E+06 | 6.91E+05 | 1.90E+04 | 2.95E+03 | 3.14E+04 | 2.10E+03 |
| $max_o$ | 2.38E+07 | 3.10E+06 | 2.27E+07 | 6.00E+06 | 1.02E+08 | 1.36E+07 | 6.49E+07 | 2.51E+07 | 2.23E+05 | 3.16E+04 | 2.16E+05 | 2.06E+04 |
| pred. values | N=163 | | N=53 | | N=163 | | N=63 | | N=163 | | N=53 | |
| $GM_p$ | 2.59E+06 | 6.64E+04 | 3.27E+06 | 2.51E+05 | 1.30E+07 | 4.86E+05 | 9.11E+06 | 4.22E+05 | 7.07E+04 | 1.95E+03 | 8.61E+04 | 3.63E+03 |
| $stddev_p$ | 1.66 | 0.05 | 1.58 | 0.08 | 1.62 | 0.04 | 1.65 | 0.07 | 1.35 | 0.02 | 1.37 | 0.03 |
| $median_p$ | 2.65E+06 | 1.55E+05 | 3.24E+06 | 3.31E+05 | 1.30E+07 | 8.22E+05 | 8.11E+06 | 3.92E+05 | 6.89E+04 | 3.01E+03 | 8.66E+04 | 4.99E+03 |
| $min_p$ | 7.91E+05 | 2.61E+04 | 1.26E+06 | 2.10E+05 | 4.07E+06 | 3.91E+05 | 3.78E+06 | 5.70E+05 | 3.89E+04 | 2.37E+03 | 4.70E+04 | 1.57E+03 |
| $max_p$ | 1.17E+07 | 3.64E+06 | 1.27E+07 | 4.97E+06 | 5.79E+07 | 1.89E+07 | 4.06E+07 | 4.96E+06 | 1.87E+05 | 2.70E+04 | 1.93E+05 | 2.09E+03 |
| MBE | 6.45E+05 | 2.37E+05 | 1.43E+06 | 6.74E+05 | 2.50E+06 | 1.00E+06 | 1.24E+06 | 1.36E+06 | 6.94E+03 | 3.30E+03 | 1.08E+04 | 7.81E+03 |
| MAE | 2.49E+06 | 2.79E+05 | 3.32E+06 | 5.99E+05 | 1.17E+07 | 9.66E+05 | 8.32E+06 | 8.10E+05 | 3.31E+04 | 2.88E+03 | 4.15E+04 | 5.80E+03 |
| MSE | 1.53E+13 | 3.25E+12 | 2.77E+13 | 1.18E+13 | 3.21E+14 | 4.17E+13 | 1.84E+14 | 6.49E+13 | 1.95E+09 | 3.83E+08 | 2.81E+09 | 7.22E+08 |
| RMSE | 3.89E+06 | 4.08E+05 | 5.14E+06 | 1.17E+06 | 1.79E+07 | 1.17E+06 | 1.34E+07 | 2.37E+06 | 4.40E+04 | 4.30E+03 | 5.26E+04 | 6.93E+03 |
| MSEs | 1.23E+13 | 3.70E+12 | 2.36E+13 | 1.14E+13 | 2.60E+14 | 5.04E+13 | 1.35E+14 | 6.33E+13 | 1.35E+09 | 3.33E+08 | 1.93E+09 | 7.26E+08 |
| MSEu | 2.94E+12 | 8.47E+11 | 4.13E+12 | 2.40E+12 | 6.15E+13 | 1.89E+13 | 4.87E+13 | 1.27E+13 | 6.05E+08 | 9.73E+07 | 8.81E+08 | 2.33E+08 |
| MSEs/ MSE | 0.80 | 0.08 | 0.83 | 0.09 | 0.80 | 0.07 | 0.71 | 0.11 | 0.68 | 0.05 | 0.67 | 0.11 |
| MSEu/ MSE | 0.20 | 0.08 | 0.17 | 0.09 | 0.20 | 0.07 | 0.29 | 0.12 | 0.32 | 0.06 | 0.33 | 0.11 |
| a | 2.91E+06 | 2.22E+05 | 3.63E+06 | 3.83E+05 | 1.43E+07 | 1.24E+06 | 1.06E+07 | 1.04E+06 | 6.57E+04 | 4.74E+03 | 8.78E+04 | 1.47E+04 |
| b | 0.02 | 0.06 | 0.01 | 0.06 | 0.02 | 0.05 | -0.01 | 0.07 | 0.11 | 0.05 | 0.03 | 0.14 |
| $r^{2*}$ | 0.01 | 0.02 | 0.01 | 0.01 | 0.01 | 0.02 | 0.01 | 0.02 | 0.03 | 0.02 | 0.03 | 0.03 |
| d | 0.32 | 0.07 | 0.34 | 0.08 | 0.32 | 0.06 | 0.28 | 0.08 | 0.46 | 0.05 | 0.43 | 0.10 |

**Table 6-1 Summary statistics for evaluation of intermediate models**

Note, $R^2$, and RSS are for the log-normalized multiple regression model, all other values apply to the linear regression of observed vs. predicted values; $r^2$, MSEs/MSE, MSEu/MSE, a, b, and d are dimensionless, all other units are $cm^{-3}$ (for ATE and AC) and $cm^{-3}\ km^{-1}$ (for ATEK); $r^{2*}$ correlation coefficient for observed vs. predicted values.

a. average total exposure

b. average total exposure per km

c. average concentration

General findings

The values for the average multiple correlation coefficients $R^2$ shown here are very similar to those calculated for the complete data set (Table 5-3) and the general patterns were confirmed, i.e. $R^2$ is higher for ATE than for ATE(W), $R^2$ is lower for ATEK than for ATEK(LED_O) and $R^2$ is lower for AC than for AC(W). This, together with the low variation in the multiple correlation coefficients shows that the method of excluding 10 % of randomly selected data records from the multiple regression calculations regularly leads to model equations similar to the ones derived for the complete data set. The *RSS* is generally higher for the all inclusive models than for the grouped data models. This indicates that the smaller data sets for grouped data have a narrower range and are thus less likely to diverge from the calculated linear regression function.

The descriptive statistics indicate that the geometric means for observed and predicted values are similar in all cases, with the predicted values slightly higher than the observed for ATE, ATEK and ATEK(LED_O) (< 6 %) and slightly lower for ATE(W), AC and AC(W) (< 12 %). The most pronounced difference was found for ATE(W) where the predicted geometric mean is approximately 11 % lower than the observed value. The calculated standard deviation values were found to be generally lower for the predicted results than for the observed values. This is due to the wider spread of values in the observed data, i.e. the average minimum of predicted values is up to three times higher and the maximum up to two times lower than the observed values. This means that the models are not able to fully reproduce the variability of the measured data.

The results for *MBE* indicate that all models tend to overpredict; ATE and ATE(W) most strongly (up to 39 %).

The correlation coefficients $r^2$ for observed vs. predicted values are very low for all models (< 0.03), suggesting that they are very inaccurate. However, the index of agreement, *d*, indicates that the models are in fact reasonably accurate, thus confirming Willmott's notion of $r^2$ being unsuitable for model evaluation purposes. The highest average *d*-values were found for AC and AC(W), 0.46 and 0.43, respectively. The values for the ATE and ATEK models are lower, ranging from 0.28 for ATEK(LED_O) to 0.34 for ATE(W).

ATE models

The evaluation statistics for both models are very similar as far as *RMSE* and *MAE*, in relation to $GM_0$, are concerned, with *RMSE* being very inflated for both models ($RMSE/GM_0$ =1.5 and 1.4 for ATE and ATE(W), respectively). The investigation of the models' fit showed a slightly higher *d*-value for the all-inclusive model ATE. This model also had a slightly better systematic/unsystematic error ratio (i.e. 80/20 compared to 83/17 for ATE(W)), which, however, is still decidedly high. Both models show a clear tendency to overpredict, ATE by 25 % and ATE(W) by 39 %.

ATEK models

The evaluation statistics show that ATEK(LED_O) performs somewhat better than ATEK, with less deviation between observed and predicted values and a more favourable systematic/unsystematic error ratio (i.e. 71/29 compared to 80/20 for ATEK). The $RMSE/GM_0$ ratio is approximately 1.5 for both models. This, together with very similar *d*-values (0.32 and 0.28), indicates that both ATEK models are very similar in terms of their predictive ability and fit. Both models show a tendency to overpredict, ATEK by 20 % and ATEK(LED_0) by 14 %.

AC models

The evaluation statistics indicate that both models perform rather similarly. They have similar systematic/unsystematic error ratios, 68/32 and 67/33 for AC and AC(W), respectively and a slight tendency to overpredict (10 % and 12 % respectively). The fact that these differences are not very pronounced and that both models also have comparable *d*-values of 0.46 and 0.43, respectively, indicates that neither of the models can be considered clearly superior.

Summary

In all three cases, the all-inclusive models performed very similarly to the associated grouped data models. This suggests that no clear advantage is gained by using the more specific models, particularly since they can only be applied for certain cases (i.e. for winter journeys or for links which are not terminated by traffic lights). It was therefore considered reasonable to disregard the grouped data models in the subsequent analysis steps and to focus only on the all-inclusive models ATE, ATEK and AC.

The statistical parameters, and in particular the *d*-value and *MBE* (showing underprediction), indicate that the two AC models perform somewhat better than the ATE and ATEK models in predicting the respective dependent variables. However, since the AC and ATEK models only provide intermediate modelling results, rather than the link exposure values modelled by ATE, the three model scenarios can only be truly compared when their ability to predict average link exposure values is investigated, as is discussed in the following section.

### 6.1.5 Final models

Using the results from the previous stage, the values predicted with the intermediate models for the three scenarios were converted to average link exposure values as specified in Phase **C** of the scenario description (Figure 6-1). This means that the results from the ATEK model were multiplied by link length and the results from the AC model by the time spent on the link (derived from link length/TRIPS speed[1]), while ATE results did not require any further conversion. For each of the scenarios, the predicted average link exposure values for the evaluation data set were then compared to the observed values in order to assess the predictive ability of the individual scenarios.

As noted above, only the all inclusive models based on the complete data set were investigated here since the grouped data models performed similar to the all-inclusive models in the previous stage of the analysis and would not have added any further detail due to the conversion process being the same for results from grouped and ungrouped data models.

In order to investigate to what extent the speed values used for the conversion of the AC results affect the predictive ability of the model, a modified version of the AC scenario was also investigated. In this version, an average speed value was calculated for each link based on the speed data collected during the field study. This speed value was then used, instead of the modelled TRIPS speed, to convert the intermediate results from the AC model into exposure values. To distinguish between the two models they are referred to as AC_T (using TRIPS data) and AC_M (using measured speed values).

1. Average speed on link during morning or evening rush hour determined by the TRIPS model (data provided by the Pollution Control Group, Leicester City Council).

The predictive ability of all four models was evaluated for exposure on individual links as well as whole journeys. The results are described in the following sections.

Predicting link exposure

Using the appropriate conversion calculations, the results from the intermediate models were converted to link exposure values (predicted) and compared with the measured data (observed) in order to assess the predictive ability of the models. The evaluation statistics for all four models are shown in Table 6-2. Similarly to the previous table, the first two rows show the average and standard deviation for $R^2$ and *RSS* for the ten regression equations calculated for each variable combination, followed by descriptive statistics of the observed and predicted total exposure values. Since all intermediate model results were based on the same 'observed values' data set and converted to 'total average exposure per link' values, the results for the different models can be used directly for cross-comparison. Although the values for $R^2$ and *RSS*, as well as the results for the ATE model, are identical to the ones in the previous table, they are included here again for comparison.

The table shows that all models requiring a conversion of intermediate results generally perform better than the ATE model which estimates average link exposure directly. All three models have a more favourable error ratio and higher d-values. They also have higher standard deviation values which indicates that they are able to better reproduce the variability inherent in the observed values.

The AC_T model has the best error ratio of all four models (36/64) and almost matches the standard deviation of the observed values. However, the model tends to overestimate exposure values by ~ 8 %, and, according to the lower *d*-value of 0.35 and inflated *RMSE* value, generally performs worse than the ATEK model. However, the results for AC_M show that the AC model can be significantly improved in some respects by using more representative speed values to convert the intermediate model results to average link exposure values.

The AC_M model has the highest *d*-value (0.71) of the investigated scenarios. Moreover, the $RMSE/GM_o$ ratio and the *MAE* value are low, and the error ratio (43/57) compares favourably to those of the ATE and ATEK models. However, the model shows less variability than the observed values and has a relatively strong tendency to underpredict, i.e. approximately 26 %.

| Model | ATE average total exposure (N=10) | | ATEK average total expos./km (N=10) | | AC_T average concentration (TRIPS speed) (N=10) | | AC_M average concentration (measured speed) (N=10) | |
|---|---|---|---|---|---|---|---|---|
| | mean | stddev | mean | stddev | mean | stddev | mean | stddev |
| $R^2$ | 0.390 | 0.003 | 0.360 | 0.004 | 0.396 | 0.003 | 0.396 | 0.003 |
| RSS | 622.87 | 8.20 | 612.12 | 8.08 | 210.23 | 1.78 | 210.23 | 1.78 |
| obs. values | N=163 | | | | | | | |
| GM | 2.58E+06 | 9.23E+04 | | | | | | |
| stddev | 2.27 | 0.13 | | | | | | |
| median | 2.56E+06 | 1.26E+05 | | | | | | |
| min | 3.04E+05 | 1.22E+05 | | | | | | |
| max | 2.38E+07 | 3.10E+06 | | | | | | |
| pred. values | N=163 | | N=163 | | N=163 | | N=163 | |
| GM | 2.59E+06 | 6.64E+04 | 2.73E+06 | 1.47E+05 | 2.30E+06 | 9.62E+04 | 3.56E+06 | 1.73E+05 |
| stddev | 1.66 | 0.05 | 1.97 | 0.11 | 2.32 | 0.11 | 1.83 | 0.04 |
| median | 2.65E+06 | 1.55E+05 | 2.72E+06 | 1.43E+05 | 2.27E+06 | 2.02E+05 | 3.51E+06 | 2.08E+05 |
| min | 7.91E+05 | 2.61E+04 | 5.01E+05 | 5.92E+04 | 3.00E+05 | 2.38E+04 | 6.44E+05 | 1.22E+05 |
| max | 1.17E+07 | 3.64E+06 | 1.69E+07 | 5.29E+06 | 3.25E+07 | 7.51E+06 | 1.73E+07 | 2.75E+06 |
| MBE | 6.45E+05 | 2.37E+05 | 1.81E+05 | 2.52E+05 | 2.07E+05 | 2.05E+05 | -6.71E+05 | 2.29E+05 |
| MAE | 2.49E+06 | 2.79E+05 | 2.54E+06 | 2.39E+05 | 2.92E+06 | 2.88E+05 | 2.11E+06 | 9.87E+04 |
| MSE | 1.53E+13 | 3.25E+12 | 1.53E+13 | 2.49E+12 | 2.71E+13 | 5.50E+12 | 9.66E+12 | 9.75E+11 |
| RMSE | 3.89E+06 | 4.08E+05 | 3.91E+06 | 3.14E+05 | 5.18E+06 | 5.21E+05 | 3.10E+06 | 1.61E+05 |
| MSEs | 1.23E+13 | 3.70E+12 | 9.23E+12 | 3.14E+12 | 9.72E+12 | 2.76E+12 | 4.19E+12 | 8.61E+11 |
| MSEu | 2.94E+12 | 8.47E+11 | 6.12E+12 | 1.89E+12 | 1.74E+13 | 3.93E+12 | 5.47E+12 | 8.42E+11 |
| MSEs/ MSE | 0.80 | 0.08 | 0.59 | 0.14 | 0.36 | 0.07 | 0.43 | 0.07 |
| MSEu/ MSE | 0.20 | 0.08 | 0.41 | 0.14 | 0.64 | 0.08 | 0.57 | 0.07 |
| a | 2.91E+06 | 2.22E+05 | 2.93E+06 | 2.85E+05 | 3.00E+06 | 3.38E+05 | 2.65E+06 | 2.45E+05 |
| b | 0.02 | 0.06 | 0.14 | 0.08 | 0.11 | 0.08 | 0.45 | 0.05 |
| $r^{2*}$ | 0.01 | 0.02 | 0.04 | 0.03 | 0.01 | 0.01 | 0.31 | 0.07 |
| d | 0.32 | 0.07 | 0.44 | 0.07 | 0.35 | 0.08 | 0.71 | 0.04 |

**Table 6-2 Summary statistics for evaluation of link exposure estimates**

Note, $R^2$, and RSS are for the log-normalized multiple regression model, all other values apply to the linear regression of observed vs. predicted values; $r^2$, MSEs/MSE, MSEu/MSE, a, b, and d are dimensionless, all other units are $cm^{-3}$ (for ATE and AC) and $cm^{-3}$ $km^{-1}$ (for ATEK); $r^{2*}$ correlation coefficient for observed vs. predicted values.

Investigating over- and underprediction

An additional visual inspection of scatterplots of the relative deviations (calculated as $Residual_i/Observed_i$) vs. observed values ($Observed_i$) revealed that the over- or underprediction indicated by *MBE* may mask a pattern of more extreme deviations. This is illustrated in Figure 6-2 for AC_T and AC_M for values from one model iteration. For both models, negative relative deviation values (i.e. underprediction) are observed throughout the exposure range and overpredictions are more likely to occur for lower average exposure values. However, for AC_T some extreme cases of positive deviation (> 1000 %) occur for low exposure values. This indicates that the low overall overprediction of the AC_T model, suggested by the reported *MBE* value, is a result of some extreme values balancing out the frequent underpredictions, while for AC_M, the underpredictions are more apparent in the *MBE* value due to less extreme individual overpredictions.

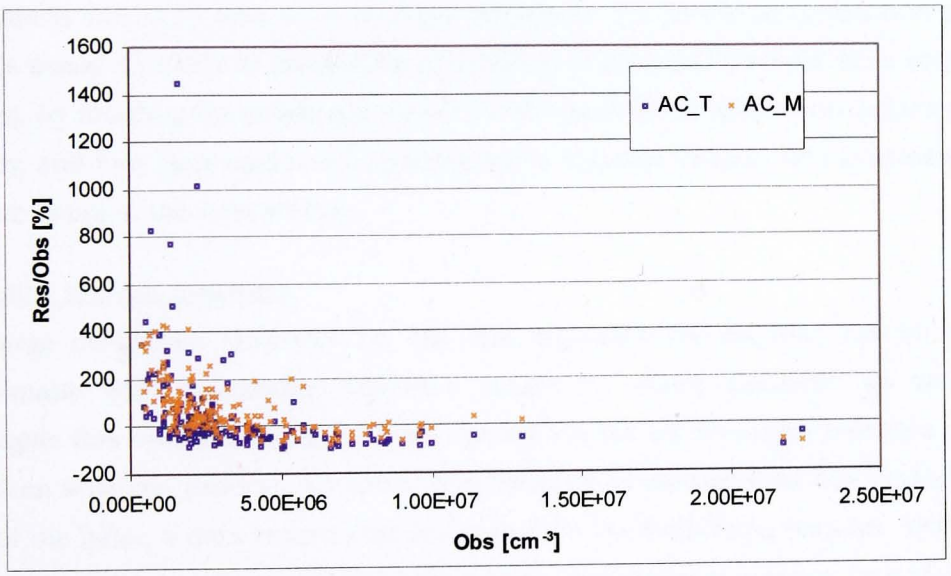


**Figure 6-2 Observed average link exposure values vs. relative deviation for results from the AC_T and AC_M models (example for one model iteration, N = 163)**

Visual inspections of relative deviations for predicted results from the ATE and ATEK models showed similar patterns to those of the AC_T model, i.e. extreme positive deviations for low exposure values. However, for ATE there were generally few cases of underprediction, resulting in the large value for *MBE*, indicating a strong general tendency to overpredict, as shown in Table 6-2.

Further investigations also showed that the extreme cases of overprediction are frequently observed for particular links. However, the types of links affected vary for the different models. For the ATE model, extreme deviations between observed and predicted values tend to be associated with very short links. For the AC models, the extreme values seem to be associated with lower speed values being used in the conversion of intermediate model results. Although no such relationship could be identified for the links associated with extreme residual values in the ATEK model, these observations indicate that the performance of the models generally declines when extreme values are used for the model input and conversion parameters. This is a typical drawback of linear regression models, which are based on the assumption of linear relationships between parameters and may thus not be able to generate accurate results for extreme values.

Although large deviations were observed for individual predictions, the statistical model evaluation parameters indicate that the models perform reasonably well, particularly the ATEK and AC_M models. Moreover, the practical application of the models would be likely to involve the calculation of an exposure total for a complete journey, by totalling the predicted values for the individual links visited during that journey, and may thus lead to a balancing out of extreme values. This is investigated in more detail in the next section.

Predicting journey exposure

The large deviations observed for the link exposure results may not be overly problematic when predicting exposure values for whole journeys. In order to investigate this issue, calculations were carried out for 10 simulated journeys. Using a random selection process, a journey consisting of 10 unique links was devised. For each of the links, a data record was selected from the evaluation data set. Using the generated model equations and the appropriate conversion processes, link exposure values were calculated. By adding up the results for the 10 links, a total journey exposure value was generated. This process, including repeated random link selection, was carried out for all ten evaluation data sets. The ten predicted journey exposure values were then compared to the observed exposure values, calculated as the sum of the observed link exposure values. Model evaluation statistics were generated for journey exposure estimates from the four models. These are shown in Table 6-3 together with descriptive statistics for observed and predicted values. In

addition, the linear relationships between observed and predicted values are shown in more detail in Figure 6-3. For each scenario, all 10 predicted journey exposure values are plotted against the associated observed values. A black line, representing the ideal case of $P_i = O_i$, has been included for reference.

| Model | ATE average total exposure | ATEK average total expos./km | AC_T average concentration (TRIPS speed) | AC_M average concentration (measured speed) |
|---|---|---|---|---|
| observed values | N=10 | | | |
| mean | 3.38E+07 | | | |
| stddev | 8.24E+06 | | | |
| median | 3.31E+07 | | | |
| min | 1.86E+07 | | | |
| max | 4.81E+07 | | | |
| predicted values | N=10 | N=10 | N=10 | N=10 |
| mean | 3.11E+07 | 3.35E+07 | 2.60E+07 | 4.12E+07 |
| stddev | 4.74E+06 | 8.32E+06 | 7.26E+06 | 8.27E+06 |
| median | 3.10E+07 | 3.24E+07 | 2.55E+07 | 3.98E+07 |
| min | 2.45E+07 | 2.30E+07 | 1.77E+07 | 3.04E+07 |
| max | 3.89E+07 | 4.61E+07 | 4.02E+07 | 5.81E+07 |
| MBE | -2.72E+06 | -2.98E+05 | -7.74E+06 | 7.40E+06 |
| MAE | 6.39E+06 | 6.64E+06 | 8.16E+06 | 7.40E+06 |
| MSE | 5.67E+13 | 5.41E+13 | 1.22E+14 | 6.70E+13 |
| RMSE | 7.53E+06 | 7.35E+06 | 1.11E+07 | 8.18E+06 |
| MSEs | 4.07E+13 | 1.15E+13 | 8.37E+13 | 5.53E+13 |
| MSEu | 1.60E+13 | 4.26E+13 | 3.88E+13 | 1.17E+13 |
| MSEs/MSE | 0.72 | 0.21 | 0.69 | 0.83 |
| MSEu/MSE | 0.28 | 0.79 | 0.32 | 0.17 |
| a | 2.22E+07 | 1.43E+07 | 1.33E+07 | 1.07E+07 |
| b | 0.26 | 0.57 | 0.38 | 0.90 |
| $r^{2*}$ | 0.21 | 0.32 | 0.18 | 0.81 |
| d | 0.65 | 0.74 | 0.60 | 0.78 |

**Table 6-3 Summary statistics for comparison of journey exposure predictions**

The descriptive statistics indicate that, on average, the predicted values from the ATEK model match the observed values better than any other model. Although the ATE model also performs reasonably well, with *MAE* and *RMSE* values similar to ATEK, its predictions fall within a rather narrow range and as indicated by the low standard deviation, the model is not able to reproduce the variability inherent in the observed values. Moreover, the error ratio of 72/28 indicates the model suffers from larger systematic errors than the ATEK model, which has an almost reversed ratio of 21/79.

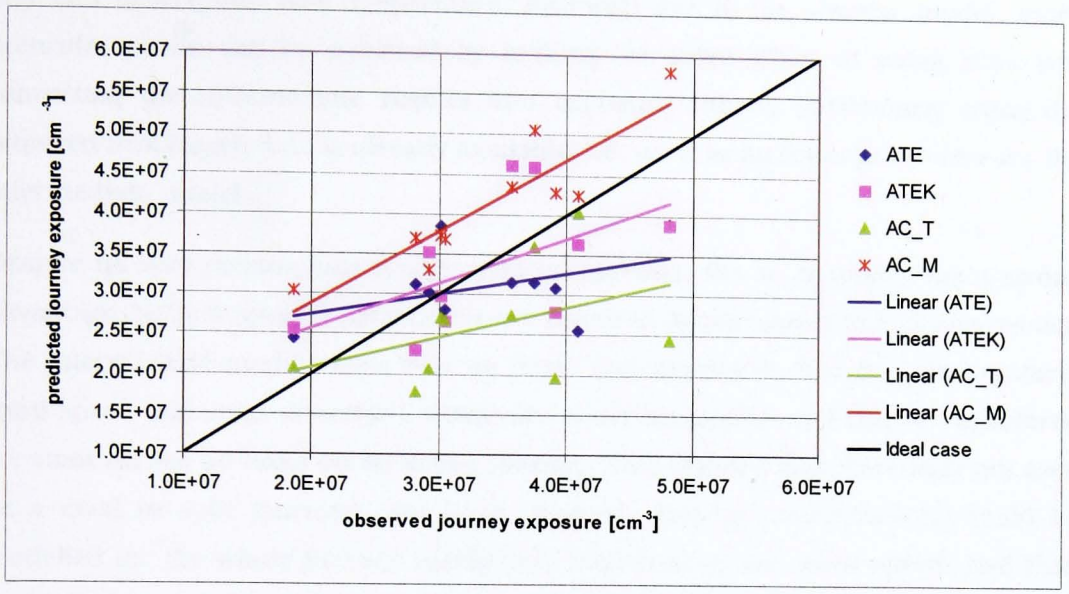


**Figure 6-3 Journey exposure, observed vs. predicted values for the four scenarios (ATE, ATEK, AC_T and AC_M)**

According to the descriptive statistics, the two AC models produce results which are further removed from the observed values, with AC_T underpredicting and AC_M overpredicting. However, as the model evaluation parameters show, AC_M is clearly the better of the two models, with lower *MAE* and *RMSE* values. In fact, AC_M has the highest $r^2$ and $d$ values of all four scenarios, indicating a good linear fit between observed and predicted values. This is particularly apparent in Figure 6-3, where the red fitted line for AC_M matches the slope of the ideal case more closely than the fitted lines for the other models. However, the fitted AC_M line does not cross the line for the ideal case, i.e. journey exposure is overestimated in all cases. Despite the larger spread of values around the fitted line, the ATEK model performs better, since over and under predictions more likely to even out.

### 6.1.6 Discussion

The analysis of the models' ability to predict personal exposure for individual links and for entire journeys has shown that the ATEK and AC_M models generally performed better than the ATE and AC_T models. However, taking additional factors such as data availability into account, all four models have their merits.

The models ATE and ATEK both require the same input parameters, i.e. link length, link end description and temperature. Although ATE is the simpler model, more accurate results can be achieved by making the extra effort of using ATEK and converting the intermediate results into exposure values, particularly since the required link length data is already available, i.e. used as an input parameter for the intermediate model.

Despite its very pronounced tendency to overpredict, the AC_M model has a strong advantage. No link specific parameters are required to calculate intermediate results. The intermediate model relies only on basic meteorological data (i.e. temperature, wind speed and wind direction), which are easily obtainable and can be considered constant across all links on an urban journey. This implies that there may not even be a need to split journeys into links. Instead, average concentrations could be modelled for the whole journey (using only temperature and wind speed) and then multiplied by the journey time. If journey time is not known, the journey length and speed data must be as accurate as possible in order to ensure reasonable results. This approach is certainly simple and results indicate that it may produce results which are only somewhat less accurate than models that take link parameters into account, such as ATEK, and thus require a more complicated set-up process. Moreover, the good linear fit observed for this model suggests that it may be possible to apply some kind of additional adjustment factor which could scale the predicted results and thus limit overprediction.

It is important to note that this analysis was based on driver exposure data from urban journeys only. It is likely that a stronger sensitivity of exposure values to link specific factors (e.g. road type, road side environment etc.) would have been observed if data from rural journeys had been included. It is possible that in that case the ATE and ATEK models may have performed better than the AC models.

This analysis has shown that it is problematic to accurately predict driver exposure based on readily available parameters such as meteorological data and link specific information. Due to the complex underlying causes of driver exposure, it is likely that better results could be achieved if more specific data was available, such as what kind of vehicle is being followed etc. However, such information would be highly specific to a particular journey and the potential for generalisation would be limited. The strong over- and underpredictions observed in some cases indicate that models derived with

multiple linear regression methods may not be ideal for the estimation of driver exposure levels. It is likely that other modelling approaches, which are not based on the assumption of linear relationships between input and output parameters, may be more suited to produce more accurate predictions, especially for particularly high or low input values.

It should be noted that the model equations which were generated as part of this analysis are not to be taken as generally applicable for driver exposure estimation. They were merely used as a means to assess the usefulness of driver exposure data for modelling purposes. In order to develop driver exposure models, additional data for a wider range of road types, weather conditions and possibly different vehicles would be required which is outside the scope of this project.

## 6.2 Annual exposure modelling

As the previous section has shown, predicting driver exposure for individual links or journeys yields mixed results. Although the data set is relatively large, the models did not produce accurate results in all cases, often resulting in large deviations between observed and predicted values. This is partly due to relatively few measurements being available for the low and high ends of the scales for the various input parameters. However, considering that exposure predictions for longer time spans, such as annual exposure, are calculated as statistical moments of numerous individual data values, extreme low or high values are more likely to cancel each other out, leading to a more general prediction which may however be reasonably accurate.

### 6.2.1 Objectives

Although the field study data may not be suitable to provide accurate short-term predictions, it may have some potential for estimating annual exposure levels. The objective of the modelling attempt described in this section was to derive a general method that would allow field study data to be utilized to estimate annual exposure averages. Although some numerical values were generated as part of this analysis, they cannot readily be used to derive general conclusions regarding long-term driver exposure due to the relatively small data set used.

It is generally accepted that particle concentrations strongly depend on weather conditions, such as temperature and wind speed (e.g. Harrison et al., 1999a; Kittelson

et al., 2000; Wehner and Wiedensohler, 2003). As discussed in the previous chapters, results from this study have shown that similar effects can be observed for driver exposure to ultrafine particles. It therefore seems reasonable that meteorological data should be used for the estimation of annual driver exposure. Accordingly, the main objective of the analysis here was to investigate whether particle exposure measurements from the field study can be used in conjunction with meteorological reference data to give an estimate of the average annual exposure experienced by individuals who commute by car. The proposed method was envisaged to be general enough to be applied to similar data sets from comparable field studies. However, since the data set used here originates from a field study carried out in an urban area, the method can in this analysis only be validated for urban data.

### 6.2.2 Methods

The proposed modelling approach is based on the concept of calculating annual exposure based on the number of occurrences of certain meteorological conditions, i.e. combinations of wind speed and temperature levels. By using the average concentration values from the field data set measured for each combination of wind speed and temperature and multiplying them by the number of occurrences of these conditions in 'a typical year' (i.e. meteorological reference data), an annual exposure estimate can be derived.

The approach required the compilation of average field study values and the extraction and preparation of the relevant meteorological reference data. This was achieved using custom written IDL routines which also facilitated suitable visual representation of the data as colour coded matrices. The individual steps of the modelling process are illustrated for ultrafine particle concentrations in Figures 6-4a-d. Average annual rush hour exposure values were also estimated for $PM_{10}$ and $PM_{2.5}$. The results for all size ranges are outlined in Section 6.2.3.

Meteorological reference data

This analysis utilizes meteorological reference data from CIBSE Test Reference Year (TRY), which is based on long-term meteorological measurements at Heathrow and Bracknell. According to the CIBSE Guide J (CIBSE, 2002), this reference data set was generated by using various statistical methods to analyse data from 20 years of measurements and compiling a composite year consisting of hourly data from the

most representative months. Originally used for dynamic thermal simulation, these data sets have recently been applied to urban solar access modelling (Mardaljevic and Rylatt, 2003).

The wind speed and temperature data from this test reference year can be used to generate a two-dimensional histogram (matrix) of temperature vs. wind speed conditions as shown in Figure 6-4a. This particular matrix was generated by extracting all hourly temperature and wind speed data for the main rush hour periods (8 - 9 am, 5 - 6 pm) from the reference year and allocating them to bins in a temperature-wind speed matrix. The colour of each bin represents the proportion of values measured during specific temperature and wind conditions, as marked on the x and y axis, respectively. Empty bins are plotted in grey. The matrix clearly shows the distribution of wind speed and temperature values, with medium values of both parameters occurring most often. Note, the highest wind speed values (9 measurements between 8 and 12 m/s) are not shown here.

Particle concentration data

In order to extract the relevant particle concentration data, the field study data was prepared as follows. Average concentration values (geometric means) were calculated for all valid data collection runs by including all data values collected during rush hour time regardless of link or route. Data was available for 116 rush hour sessions, 53 from morning and 63 from evening sessions, lasting from 10 to 60 minutes. By allocating these average values to bins depending on meteorological conditions present during each session, and calculating the proportion of values present in each bin, a matrix was generated as shown in Figure 6-4b. The total of all bin proportions equals 1.

In the next step, average field study concentrations were calculated. The geometric mean was determined for the values in each bin of the field study matrix, and a further colour-coded plot was generated as shown Figure 6-4d. Here, the colour of the bin indicates the average particle concentration measured during particular wind speed and temperature conditions. The matrix shows that, generally, higher concentrations were measured for lower temperatures and wind speeds, which is in accordance with results reported above and findings from other studies. For further details refer to the discussion in Section 4.2.

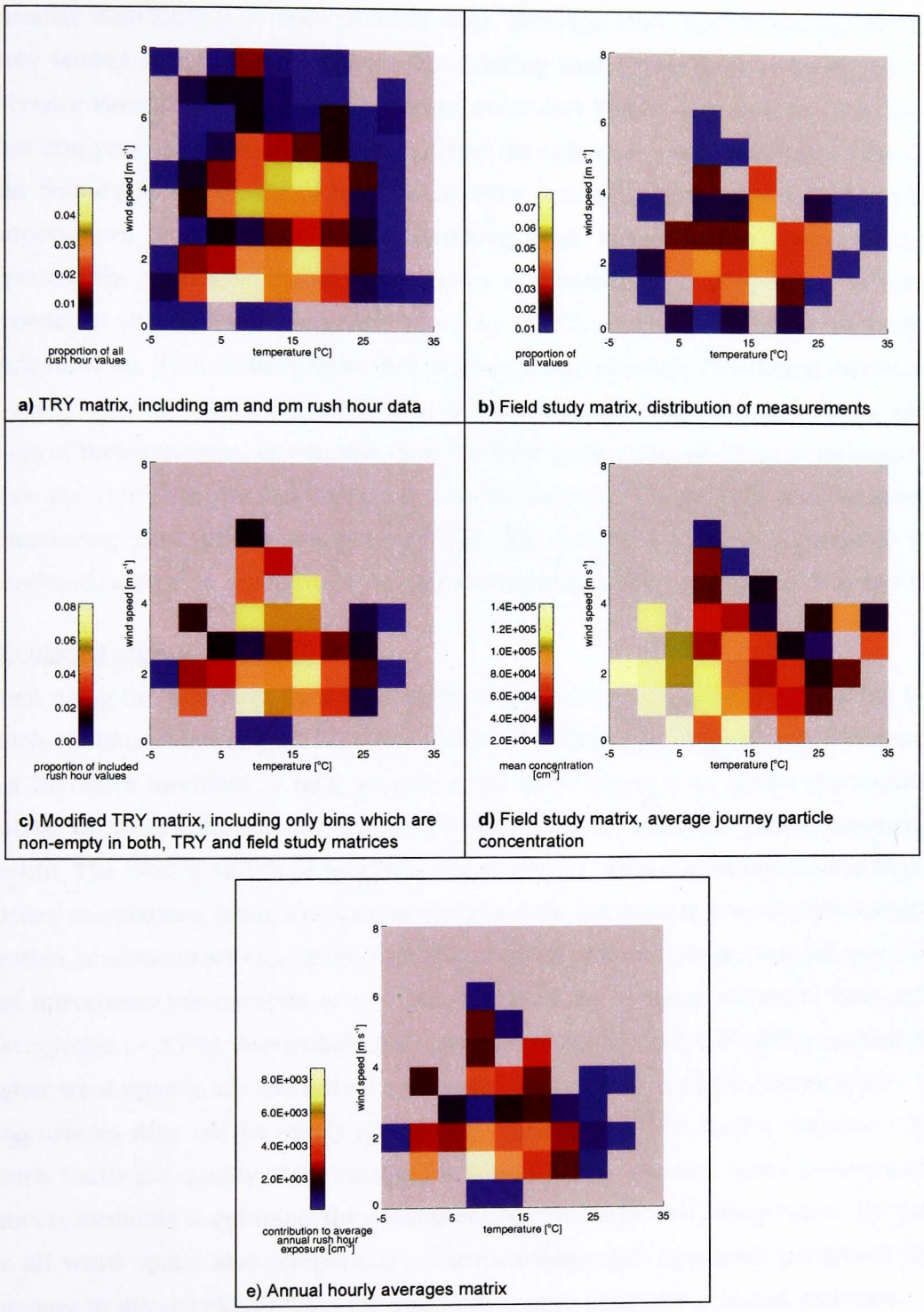


**Figure 6-4 Temperature vs. wind speed matrices illustrating the steps required to generate the annual exposure matrix form field study and reference data**

Comparison of the data sets

Comparison with the meteorological reference data shows that the field study data has a similar distribution as the reference data, although wind speeds during the field study tended to be lower. However, considering that Figure 6-4a is based on 721 reference values from long-term measurements and Figure 6-4b only on 116 values from one year, the field study data matches the reference year reasonably well. The light colours in the centre of the matrix show that most data were collected when temperatures were around 15 °C, matching peak values in the reference data. However, the reference matrix also shows a relatively high proportion of values recorded at slightly lower temperatures (~ 5 - 10 °C), which are lacking from the field study data set. This is likely to be due to fewer data collection runs being carried out in winter than in summer (Section 3.1.7). The reference data clearly covers a wider range of meteorological conditions than the field study data, which is to be expected since the reference data is based on a large number of long-term measurements. Considering these differences between the data sets, they are still reasonably well correlated, with $r^2 = 0.33$ (based on number distributions, i.e. Figures 6-4a and b).

Calculating annual exposure

When using the reference data to calculate annual exposure estimates, only the bins which contain values in both matrices will be included. The original reference matrix was therefore modified to only include these bins. Figure 6-4c shows the modified matrix, with bin colour indicating the proportion of all included values allocated to the bin. The total of all bin proportions thus equals 1. This conversion means that all further calculations using this matrix are based on the assumption that the included weather conditions are representative of a complete annual data set. It is accepted that this introduces inaccuracies since a large part of the original reference data set is disregarded (~ 37%), particularly for higher wind speed values. However, noting that higher wind speeds are associated with lower concentration and exposure levels, the inaccuracies may not be overly significant. Nevertheless, the lack of exposure data clearly limits the quality of the results for this analysis. However, since no alternative data are available to complete the data set (i.e. so that exposure values can be included for all wind speed and temperature combinations), the approach presented here manages to nevertheless utilize the available data to calculate a 'broad' estimate.

In the final step, each bin value in the reference matrix was multiplied with the corresponding bin of the average concentration matrix. The resulting matrix is shown in Figure 6-4e. Due to the nature of the modified reference matrix, this process weighs the average concentrations measured for each temperature/wind speed bin by the proportion of occurrences, i.e. values < 1. An annual average for rush hour concentrations of particles can thus be derived from the result matrix by summing the values in all bins.

### 6.2.3 Results and discussion

Exposure estimates

The described method was applied to field study data for ultrafine particles as well as $PM_{10}$ and $PM_{2.5}$ mass concentrations. The results for all three size ranges are presented in Table 6-4. The annual average for exposure to ultrafine particulate matter during rush hour, calculated with the described method, is 76,950 $cm^{-3}$. Estimated average rush hour concentrations for $PM_{10}$ and $PM_{2.5}$ are 48.53 $\mu g\, m^{-3}$ and 18.48 $\mu g\, m^{-3}$, respectively.

Interpolation of missing data

Since the incomplete coverage of field study data is thought to affect the accuracy of these results, it was investigated whether standard interpolation methods could be used to generate values for some of the empty bins of the matrix and thus improve the results. The interpolation method and the derived matrices are described in Appendix 6. Although the correlation between the reference data and the number distribution matrix for the field data could be increased to $r^2 = 0.45$ by including interpolated values, the estimated rush hour averages based on this more complete matrix differed only slightly from the results derived with the original matrix (Table 6-4). However, the interpolation method allows the calculations to be based on more meteorological reference data and may thus be useful for the estimation of annual exposure from less comprehensive data sets.

Comparison of modelled and measured values

As shown in Table 6-4, the estimated average annual exposure values differ little from the average exposure calculated for the field study measurements.

This is thought to be due to the field study including data from all seasons, i.e. for a wide range of weather conditions. As explained above, the number distribution matrices for meteorological data from the field study match the reference data well. In cases where the measurements are carried out during less representative meteorological conditions, the estimation method proposed here is likely to provide a better estimate of average annual exposure than the average of those measurements, and a more pronounced difference would be observed between the two values.

| Particle size range | Average measured journey exposure | Estimated average annual rush hour exposure | Estimated average annual rush hour exposure using interpolated values | Air quality standard |
|---|---|---|---|---|
| UF | 76,652 $cm^{-3}$ | 76,950 $cm^{-3}$ | 77,140 $cm^{-3}$ | - |
| $PM_{10}$ | 52.16 $\mu g\ m^{-3}$ | 48.53 $\mu g\ m^{-3}$ | 46.98 $\mu g\ m^{-3}$ | 40 $\mu g\ m^{-3}$ (DEFRA, 2003a) |
| $PM_{2.5}$ | 20.19 $\mu g\ m^{-3}$ | 18.48 $\mu g\ m^{-3}$ | 18.41 $\mu g\ m^{-3}$ | 15 $\mu g\ m^{-3}$ (USEPA, 1997) |

**Table 6-4 Comparison of modelled and measured exposure parameters for ultrafine particles (UF), $PM_{2.5}$ and $PM_{10}$**

It is important to note that the calculated estimates for average annual rush hour exposure are not suitable for estimating the exposure during individual journeys, but rather provide a factor for estimating the annual exposure based on the number of hours spent commuting in a year along the field study routes or comparable areas. For example, using the derived estimation factor and assuming that there are approximately 220 working days in a year on which a commuter may spend two hours in urban traffic, one hour each during morning and evening rush hour, the annual integrated rush hour exposure to ultrafine particulate matter experienced by the commuter would be expected to be approximately 34 $x10^6$ $cm^{-3}$ h.

Comparison with air quality standards and fixed site monitoring data

However, as noted in Section 2.3, exposure values are more commonly presented as 'average exposure', giving the average concentration observed during the time period of interest. The derived exposure estimates can therefore be compared to other relevant annual average concentrations (or exposures) such as air quality standards. However, no air quality standards have been set for ultrafine particles (UF), which could be used to qualify these results. However, there is a standard in force in the UK for $PM_{10}$ particles and the USEPA[2] have set a standard for $PM_{2.5}$ particles. Both

2. US Environmental Protection Agency.

values are included in Table 6-4. As outlined in Section 2.1, these standards are objectives for annual average concentrations, introduced for air quality assessment purposes. For both particle size ranges, the estimated rush hour exposure is higher than the air quality standard, which is to be expected since the estimates are based on rush hour measurements only while annual averages, which are assessed using the standards, also incorporate data from off-peak and night time periods when concentrations tend to be significantly lower.

It therefore seems more appropriate to compare the driver exposure estimates to monitored rush hour data. For this purpose, relevant particle concentration data sets were acquired from the UK National Air Quality Information Archive[3]. The data are provided as hourly averages measured at roadside and background locations. Only data for $PM_{10}$ and $PM_{2.5}$ were available. By using hourly averages measured during morning and evening rush hour (i.e. 8 - 9 am and 5 - 6 pm) at automatic monitoring stations, annual average values for rush hour periods were calculated. Figure 6-5 shows the estimated driver exposure values together with the respective air quality standards and annual rush hour averages from selected monitoring stations for both, $PM_{10}$ and $PM_{2.5}$ size ranges. These values were calculated based on hourly average values measured during morning and evening rush hour at the monitoring stations throughout the field study period, i.e. 01/05/02 to 30/04/03. It is thought that by using fixed site monitoring data from the twelve months that exactly match the field study period, possible elevated pollution levels due to long-range transport of particulate matter from mainland Europe may be indirectly accounted for.

The monitoring stations from which data was used were selected based on the following considerations[4]. 'Leicester Centre' (urban centre) and 'Nottingham Centre' (urban centre) were selected due to their relative proximity to the field study area. 'London Marylebone Road' (kerbside) is included due to being the station with the highest pollution levels in the UK (APEG, 1999) and 'London N. Kensington' (urban background) as a background reference for London. For $PM_{2.5}$, the four stations included are the only ones on the automatic monitoring network that routinely measure $PM_{2.5}$ concentrations.

3. Accessible at http://www.airquality.co.uk.
4. The classification of monitoring stations given in brackets for each site is based on those used by the AUN site information archive (http://www.stanger.co.uk/siteinfo/).

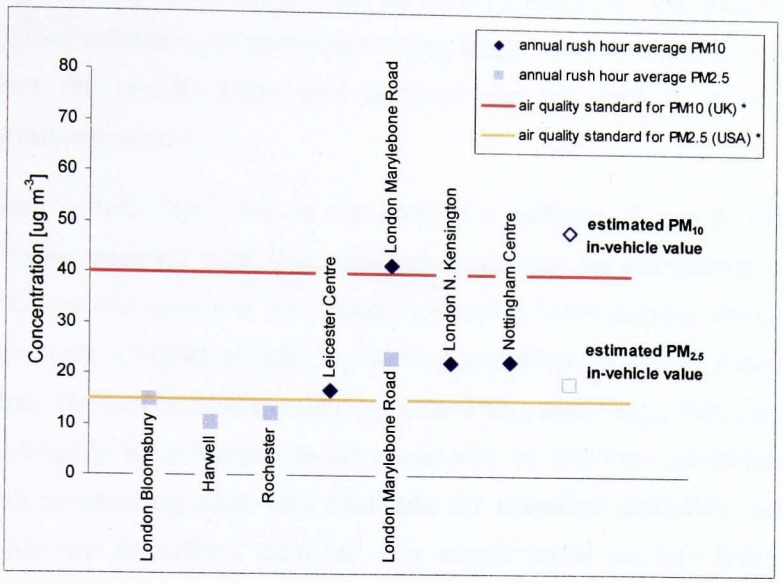


**Figure 6-5 Comparison of estimated annual exposure levels with measured annual averages and air quality standards (* objectives for average annual concentration)**

Figure 6-5 shows that the average values for $PM_{10}$ from the monitoring stations are generally lower than the driver exposure estimate, particularly the result for Leicester City Centre. Only the rush hour average for London Marylebone Road, a kerbside monitoring station where particularly high pollution levels are frequently observed (APEG, 1999), reaches similar levels to the driver exposure estimate. Similar observations can be made for concentrations of particles in the $PM_{2.5}$ range. Here the value for London Marylebone Road is even higher than the estimate for annual average driver exposure.

### 6.2.4 Conclusions

These comparisons seem to indicate that average annual driver exposure to $PM_{10}$ and $PM_{2.5}$ particles exceeds average exposure to outside concentrations during rush hour. This is in accordance with results from a number of studies that compared in-vehicle particle exposure to outside concentrations during commuter journeys (e.g. Morton, 2002; Gee et al., 1999; Adams et al., 2001b; Kingham et al., 1998).

However, as noted above, this analysis did not aim to provide general conclusions on annual driver exposure, but rather to develop a method that utilizes field study data to make an estimation of annual average values. It is therefore interesting to note that

the estimated parameters are indeed comparable to rush hour monitoring data for both size ranges, $PM_{10}$ and $PM_{2.5}$. This agreement supports the proposed method. However, a formal validation is necessary, using larger data sets and data from various studies, before the results from this method can be used to draw conclusions regarding annual exposure.

It is important to note that, due to the lack of a suitable data set, the results for ultrafine particles derived with this method could not be compared with external measurements. As the analyses in Chapters 4 and 5 have shown, concentrations of these particles have a higher spatial variability and show different patterns than the larger particles. The observations made here for $PM_{10}$ and $PM_{2.5}$ (e.g. Figure 6-5), can therefore not readily be assumed to be applicable to ultrafine particles also. Thus, even if annual monitoring data was available for ultrafine particles, average values calculated with the described method may significantly deviate from monitoring averages at different locations.

The results presented here suggest that personal exposure data from microenvironment field study campaigns can be used in conjunction with meteorological reference data to estimate annual exposure to $PM_{10}$ and $PM_{2.5}$. It is not clear whether the proposed method can also be successfully applied to ultrafine particles. Further validation with more comprehensive data sets (i.e. from various field studies and locations) is required in order to fully determine the method's accuracy and limitations.

## 6.3 Summary

In this chapter two different approaches for the use of field study data for driver exposure modelling were explored. The modelling methods used for both journey and annual exposure estimation were designed to ensure a commensurate level of detail at all stages of the process beginning with the data collection. This was based on the consideration that no advantage can be gained by expending disproportionate effort at one stage if uncertainties in later stages are likely to 'wash out' the additional level of detail.

The results presented here indicate that it is problematic to accurately estimate driver exposure for individual road links and journeys using the proposed method which is based on multiple linear regression models. It is possible that better results could be

achieved by using alternative modelling methods, i.e. ones that do not rely on linear relationships between input and output parameters. However, a good model fit, despite consistent overprediction, was found for the AC_M model. In order to achieve reliable estimates with this method, accurate speed values would be required and a correction factor would have to be established. Nevertheless, the good results derived with this model suggest that it may be possible to estimate journey exposure based on meteorological parameters only, without the inclusion of link specific parameters.

Taking this idea one step further, a method was proposed in the second part of this chapter which used field study data in conjunction with meteorological reference data to derive an average annual exposure estimate. The results indicate that the method produces estimates which are comparable to measured outside concentrations. However, validation is required to confirm the method's suitability for exposure estimation.

It is important to note, however, that the findings from this analysis are based on a specific field data set from urban journeys and are thus only indicative for urban traffic scenarios similar to those encountered in Leicester.

It can be concluded that empirical data from personal exposure studies can, in principle, be used to estimate short-term and long-term driver exposure, but further research and validation is required in order to develop models for practical application.

Chapter

# 7 Discussion and Conclusions

## 7.1 Summary

The aim of this study was to investigate the exposure of humans to particulate matter in a transport microenvironment. Motivated by findings from health studies and personal exposure research, the study aimed to contribute to the understanding of personal exposure to ultrafine particles in vehicles. A field study campaign was designed and carried out, which provided a unique data set of high-frequency personal exposure measurements (ultrafine particles, $PM_{10}$, $PM_{2.5}$), and related road and event data. Based on the field study data set, an analysis of the dynamics and the main determinants of personal exposure to particulate matter was carried out. The potential of the field data for exposure modelling was also explored.

First, a review was carried out of existing legislation and research on particulate matter, regarding physical and chemical composition and processes, health effects and traffic related characteristics. The theoretical background of personal exposure research as well as previous personal exposure studies in transport microenvironments were also studied (Chapter 2).

Based on findings from this review, a field study methodology was developed and applied during an extensive field campaign (Chapter 3). An instrumented vehicle was equipped with environmental monitoring devices and used to collect in-vehicle particle concentration data on three pre-defined routes. The acquired data set was processed and analysed using a set of purpose written programming routines.

During the first part of the analysis, a number of methods were applied in order to explore the dynamics of driver exposure to particulate matter (Chapter 4). The short-term variability as well as summary metrics were investigated for the different particle size fractions. This was followed by a multiple linear regression analysis which aimed to identify the main determinants, considering both meteorological and road specific parameters (Chapter 5).

In Chapter 6, methods of estimating driver exposure to particulate matter were explored. Three scenarios were proposed for the estimation of link and journey exposure, using the field study data and findings from the previous steps of the data analysis. The scenarios were compared in terms of their accuracy, and advantages and short-comings were discussed. Finally, a method was proposed and investigated which aimed at utilizing the field study data for annual exposure estimation.

## 7.2 Conclusions

The findings from this study can be summarized as follows:

- In-vehicle particle concentrations of all investigated size ranges (ultrafine, $PM_{10}$, $PM_{2.5}$) generally follow the patterns found in other studies for outside concentrations.
- In-vehicle concentrations are affected by seasonal and diurnal variations of meteorological conditions and the route travelled. Elevated concentrations of ultrafine particles are generally observed during the winter months and in the mornings, and on routes with main roads (i.e. dual carriageways).
- There is evidence that extreme exposure events can often be associated with individual highly polluting vehicles in the traffic stream.
- Results indicate that, on average, approximately 50 % of the driver exposure to ultrafine particles on urban journeys can be attributed to elevated concentrations from distinct transient traffic events.
- Mass and number concentrations show different short-term variability and sensitivity to traffic events. Peaks in number concentration measurements are often not matched in the mass concentration traces.

- The analysis of the main determinants of driver exposure showed that queuing (here represented by traffic light arrangement and link length) affects link exposure values most strongly.

- Together with season and wind conditions (speed and direction), queuing explains approximately one third of the variability in ultrafine particle exposure values.

- The main determinants of link exposure to mass concentrations ($PM_{10}$, $PM_{2.5}$) differ from those for number concentrations (ultrafine particles), i.e. they are strongly related to $PM_{10}$ background concentrations, measured at fixed site monitoring station, rather than to season and wind speed.

- Due to the diversity of conditions, it is difficult to accurately predict journey exposure based on readily available parameters. Linear regression modelling may not be an appropriate method for this purpose since extreme over- and underpredictions were observed for particularly low values of the input parameters.

- It may be possible to use exposure measurements from field studies to estimate long-term exposure averages. The method proposed here for this purpose produces results which lie in a similar range as data measured at fixed site monitors. However, additional data are required to validate the method.

The analysis has shown that route specific parameters, such as queuing at traffic lights, strongly affect driver exposure values, often more strongly than meteorological conditions. This suggests that changes in the road layout which reduce queuing may result in significant reductions of the exposure to particulate matter experienced by car commuters. However, behavioural changes may be effective also, i.e. avoiding the rush hour traffic.

It is important to note that the findings presented in this thesis apply to personal exposure to particulate matter on urban road networks only. It is likely that the inclusion of data from rural roads would have had a significant influence on the results. For example, the variability in the data set would probably be greater, due to lower in-vehicle concentrations in rural areas which are typically less polluted. Therefore, the results from the general analysis (Chapter 4) regarding total journey exposure values and the contribution of peaks to overall exposure would have been

affected, with lower minimum exposure values expected for rural journeys. Moreover, it is likely that a stronger sensitivity to road specific parameters would have been identified in the multiple regression analysis (Chapter 5). However, it is not clear how the inclusion of data from rural journeys would have affected the results of the modelling exercises (Chapter 6).

It is thus acknowledged that the findings presented here may have limited potential for extrapolation in terms of driver exposure in general due to the design of the data collection methodology used. Particularly the focus on a small number of central urban roads, their division into relatively short road links with different driving conditions and the choice of specific monitoring periods with high traffic flows restricts the interpretation of results in a wider context. It is likely that road specific parameters would be found to have a stronger influence on driver exposure values if a wider range of road and traffic conditions were to be considered.

Nevertheless, the findings from this study strongly indicate that there are significant differences between the exposure characteristics for ultrafine particles and those for $PM_{10}$ and $PM_{2.5}$, in terms of both average values and short-term variability. This further supports the hypothesis that mass concentration measurements cannot be used as a surrogate for ultrafine particles. More importantly, the findings from this study show that this hypothesis, which had been developed for outside concentrations, also applies to the microenvironment 'in-vehicle'. In order therefore to assess driver exposure to particulate matter, both mass and number concentration have to be taken into account, especially considering that the smaller particles are suspected to be more dangerous to human health.

## 7.3 Suggestions for further work

The field study methodology was developed particularly for this study. However, due to practical limitations, the ideal study implementation could not be achieved. For further field studies it would be advantageous to devise a fully integrated monitoring unit, perhaps using GPS to pinpoint the position on route. This would allow the selection of road sections of interest (i.e. defining of links) even after the completion of the field study, since positioning data would be included with the measurements. Other considerations are the inclusion of voice recordings with automatic time stamp allocation, integrated with the measured data, and logging of data to an on-board

computer. These features would minimize the processing time required for each data collection run and would ensure a more accurate time and location specification than the stopwatch approach used in this study.

The investigation of time series plots in conjunction with event data has indicated that emissions from the vehicle directly in front of the instrumented vehicle have a strong influence on in-vehicle concentrations. The event data obtained during the field campaign could be used to further investigate this issue.

It would be informative to apply the proposed model scenarios and equations for journey exposure modelling to other data sets in order to validate their performance. However, since they are quite specific, the data sets would have to be similar to the one obtained from the field study. Moreover, the results presented here indicate that linear regression models may not be suitable for this type of journey exposure modelling. It would therefore be advisable to investigate alternative modelling methods.

The modelling method proposed for the estimation of average annual exposure should be validated with data from other studies. An investigation of alternative interpolation method for filling in missing data may help to improve the accuracy of the predictions.

This study has investigated driver exposure to particulate matter during urban journeys. However, in order to gain a more complete understanding of the variability of driver exposure for different environments it would be informative carry out a similar investigation for rural journeys.

# *Bibliography*


Adams, H. S., Nieuwenhuijsen, M. J., Colvile, R. N., McMullen, M. A. S. and Khandelwal, P., 2001a, *Fine particle ($PM_{2.5}$) personal exposure levels in transport microenvironments, London, UK*, Science of the Total Environment, Vol. 279 (1-3), pp. 29-44

Adams, H. S., Nieuwenhuijsen, M. J. and Colvile, R. N., 2001b, *Determinants of fine particle ($PM_{2.5}$) personal exposure levels in transport microenvironments, London, UK*, Atmospheric Environment, Vol. 35, pp. 4557-4566

Adams, H. S., Nieuwenhuijsen, M. J., Colvile, R. N., Older, M. J. and Kendall, M., 2002, *Assessment of road users' elemental carbon personal exposure levels, London, UK*, Atmospheric Environment, Vol. 36, pp. 5335-5342

AEAT, 1998, *Predicting $PM_{10}$ concentrations in the UK*, report produced for the DETR, Report number: AEAT-4630, AEA Technology, National Environmental Technology Centre, Abingdon, UK

AEAT, 2001, *UG219 TRAMAQ - Cold start emissions. Summary report*, Report number: AEAT/ENV/R/0638, AEA Technology, Engines and emissions, Didcot, UK

AEAT, 2004, *Local Authority Air Pollution Monitoring Helpline: Operational Report for January to December 2003*, report produced for DEFRA, Report number: AEAT/ENV/R/1699, AEA Technology, Abingdon, UK

Allison, P. D., 1999, *Multiple regression; A primer*, Pine Forge Press, Thousand Oaks, California, USA

Alm S., Jantunen M. J. and Vartiainen, M., 1999, *Urban commuter exposure to particle matter and carbon monoxide in an automobile*, Journal of Exposure Analysis and Environmental Epidemiology, Vol. 9, pp. 237-244

APEG, 1999, *Source apportionment of airborne particulate matter in the United Kingdom*, Report of the Airborne Particles Expert Group, Department of the Environment, Transport and the Regions, UK

AQEG, 2002, *Minutes–Meeting of the 19th September 2002*, Air Quality Expert Group, Department for Environment, Food & Rural Affairs, UK, http://www.defra.gov.uk/environment/airquality/aqeg/meetings/020919min.pdf, accessed June 2004

Arnold, S., ApSimon, H., Barlow, J., Belcher, S., Bell, M., Boddy, D., Britter, R., Cheng, H., Clark, R., Colvile, R., Dimitroulopoulou, S., Dobre, A., Greally, B., Kaur, S., Knights, A., Lawton, T., Makepeace, A., Martin, D., Neophytou, M., Neville, S., Nieuwenhuijsen, M., Nickless, G., Price, C., Robins, A., Shallcross, D., Simmonds, P., Smalley, R., Tate, J., Tomlin, A., Wang, H. and Walsh, P., 2004, *Dispersion of air pollution and penetration into the local environment DAPPLE*, Science of the Total Environment, In Press

Ayers, G. P., Keywood, M. D. and Gras, J. L., 1999, *TEOM vs. manual gravimetric methods for determination of $PM_{2.5}$ aerosol mass concentrations*, Atmospheric Environment, Vol. 33, pp. 3717-3721

Ayers, G. P., 2001, *Comment on regression analysis of air quality data*, Technical note, Atmospheric Environment, Vol. 35, pp. 2423-2425

Baker, C. J., 2001, *Flow and dispersion in ground vehicle wakes*, Journal of Fluids and Structures, Journal of Fluids and Structures, Vol. 15(7), pp. 1031-1060

Barratt, R., 2001, *Atmospheric dispersion modelling, An introduction to practical applications*, Earthscan Publications Ltd., London, UK

Behrentz, E., Fitz, D. R., Pankratz, D. V., Sabina, L. D., Colomec, S. D., Fruind, S. A., Winera, A. M., 2004, *Measuring self-pollution in school buses using a tracer gas technique*, Atmospheric Environment, Vol. 38, pp. 3735-3746

Bukowiecki, N., Dommen, J., Prévôt, A. S. H., Weingartner, E. and Baltensperger, U., 2003, *Fine and ultrafine particles in the Zuerich (Switzerland) area measured with a mobile laboratory. An assessment of the seasonal and regional variation throughout a year*, Atmospheric Chemistry and Physics Discussions, Vol. 3, pp. 2739–2782

Bukowiecki, N., Dommen, J., Prévôt, A. S. H., R. Richter, Weingartner, E. and Baltensperger, U., 2002, *A mobile pollutant measurement laboratory - measuring gas phase and aerosol ambient concentrations with high spatial and temporal resolution*, Atmospheric Environment, Vol. 36, pp. 5569–5579

Census, 2001, Key Statistics, Crown Copyright, Leicester City Council Area Profile Version 1.1, http://www.leicester.gov.uk/departments/print.asp?pgid=1009, accessed June 2004

Chaloulakou, A., Grivas, G. and Spyrellis, N., 2003, *Neural network and multiple regression models for $PM_{10}$ prediction in Athens: A comparative assessment*, Journal of the Air & Waste Management Association, Vol. 53, pp. 1183-1190

Chan, Y. C., Simpson, R. W., McTainsh, G. H., Vowles, P. D., Cohen, D. D. and Bailey, G. M., 1997, *Characterisation of chemical species in $PM_{10}$ and $PM_{2.5}$ aerosols in Brisbane, Australia*, Atmospheric Environment, Vol. 31, pp. 3773-3785

Chan, A. T. and Chung, M. W., 2003, *Indoor-outdoor air quality relationships in vehicle: Effect of driving environment and ventilation modes*, Atmospheric Environment, Vol. 37, pp. 3795-3808

Chan, C.-C., Chuang, K.-J., Shiao, G.-M. and Lin, L.-Y., 2004, *Personal exposure to submicrometer particles and heart rate variability in human subjects*, Environmental Health Perspectives, Vol. 112, pp. 1063-1067

Chang, L., Koutrakis, P., Catalano, P. J. and Suh, H. H., 2000, *Hourly personal exposure to fine particles and gaseous pollutants - Results from Baltimore, Maryland*, Journal of the Air & Waste Management Association, Vol. 50, pp. 1223-1235

Charron, A. and Harrison, R. M., 2003, *Primary particle formation from vehicle emissions during exhaust dilution in the roadside atmosphere*, Atmospheric Environment, Vol. 37, pp. 4109-4119

CIBSE, 2002, *CIBSE Guide J: Weather, solar and illuminance data*, Department of Trade and Industry, The Chartered Institution of Building Services Engineers, CIBSE, London, UK

Colvile, R. N., Hutchinson, E. J., Mindell, J. S. and Warren, R. F., 2001, *The transport sector as a source of air pollution; Millennial review*, Atmospheric Environment, Vol. 35, pp. 1537-1565

COMEAP, 1995, *Non-biological particles and health, Executive summary*, Committee on the Medical Effects of Air Pollutants, Department of Health, UK, http://www.advisorybodies.doh.gov.uk/comeap/statementsreports/airpol1.htm, accessed June 2004

COMEAP, 1998, *Statement on the banding of air quality*, Committee on the Medical Effects of Air Pollutants, Department of Health, UK, http://www.advisorybodies.doh.gov.uk/comeap/statementsreports/airpol9.htm, accessed June 2004

COMEAP, 2001, *Statement on long-term effects of particles on mortality*, Committee on the Medical Effects of Air Pollutants, Department of Health, UK, http://www.advisorybodies.doh.gov.uk/comeap/statementsreports/longtermeffects.pdf, accessed June 2004

Deacon, A. R., Derwent, R. G., Harrison, R. M., Middleton, D. R. and Moorcroft, S., 1997, *Analysis and interpolation of measurements of suspended particulate matter at urban background sites in the United Kingdom*, Science of the Total Environment, Vol. 203 (1), pp. 17-36

DEFRA, 2000, Chapter 8: Review and assessment of $PM_{10}$ in *Review and assessment: Pollutant specific guidance*, Department for Environment, Food & Rural Affairs, UK

DEFRA, 2003a, *The Air Quality Strategy for England, Scotland, Wales and Northern Ireland: Addendum*, Department for Environment, Food & Rural Affairs, UK

DEFRA, 2003b, *Local Air Quality Management, Technical Guidance, LAQM.TG(03)*, Part IV of the Environment Act 1995, Department for Environment, Food & Rural Affairs, UK

DETR, 2000a, *The Air Quality Strategy for England, Scotland, Wales and Northern Ireland, Working Together for Clean Air*, Department of the Environment, Transport and the Regions, UK

DETR, 2000b, *Local Air Quality Management Guidance, LAQM.G1(00) - Framework for review and assessment of air quality*, Part IV of the Environment Act 1995, Department of the Environment, Transport and the Regions, UK

DETR, 2000c, *Local Air Quality Management Guidance, LAQM.G3(00) - Air Quality and Transport*, Part IV of the Environment Act 1995, Department of the Environment, Transport and the Regions, UK

DETR, 2000d, *Local Air Quality Management, Technical Guidance, LAQM.TG1(00) - Review and assessment: Monitoring air quality*, Part IV of the Environment Act 1995, Department of the Environment, Transport and the Regions, UK

DfT, 2002, *The Transport Statistics Bulletin, Traffic speeds in English urban areas: 2002*, Department for Transport, London, UK

Dickens, C. J., 2000, *In-car particle exposure*, Report produced for DETR (AEAT/ EEQA-0125), AEA Technology, Abingdon, UK

EHHI, 2002, *Children's exposure to diesel exhaust on school buses*, Environment and Health, Inc., North Haven, Connecticut, USA

Ebelt, S., Petkau, A. J., Vedal, S., Fisher, T. V. and Brauer, M., 2000, *Exposure of chronic obstructive pulmonary disease patients to particular matter: Relationships between personal and ambient concentrations*, Journal of the Air & Waste Management Association, Vol. 50, pp. 1081-1094

Ericsson, E., 2001, *Independent driving pattern factors and their influence on fuel-use and exhaust emission factors*, Transportation Research Part D, Vol. 6 (5), pp. 325-345

Everitt, B. S. and Dunn, G., 2001, *Applied multivariate data analysis*, Arnold Publishers, London, UK

Gómez-Perales, J. E., Colvile, R. N., Nieuwenhuijsen, M. J., Fernández-Bremauntz, A., Gutiérrez-Avedoy, V. J., Páramo-Figueroa, V. H., Blanco-Jiménez, S., Bueno-López, E., Mandujano, F., Bernabé-Cabanillas, R. and Ortiz-Segovia, E., 2004, *Commuters' exposure to $PM_{2.5}$, CO, and benzene in public transport in the metropolitan area of Mexico City*, Atmospheric Environment, Vol. 38, pp. 1219-1229

Gulliver, J. and Briggs, D. J., 2004, *Personal exposure to particulate air pollution in transport microenvironments*, Atmospheric Environment, Vol. 38, pp. 1-8

Harrison, R. M., Deacon, A. K., Jones, M. R. and Appleby, R.S., 1997, *Sources and processes affecting concentrations of $PM_{10}$ and $PM_{2.5}$ particulate matter in Birmingham (UK)*, Atmospheric Environment, Vol. 31, pp 4103-4118

Harrison, R. M., Yin, J., Mark, D., Stedman, J., Appleby, R. S., Booker, J. and Moorcroft, S., 2001, *Studies of the coarse particle (2.5-10um) component in UK urban atmospheres*, Atmospheric Environment, Vol. 35, pp. 3667-3679

Harrison, R. M., Jones, M. and Collins, G., 1999a, *Measurements of the physical properties of particles in the urban atmosphere*, Atmospheric Environment, Vol. 33, pp. 309-321

Harrison, R. M., Shi, J. P. and Jones, M., 1999b, *Continuous measurements of aerosol physical properties in the urban atmosphere*, Atmospheric Environment, Vol. 33, pp. 1037-1047

Hertel, O., De Leeuw, F. A. A. M., Raaschou-Nielsen, O., Jensen, S. S., Gee, D., Herbarth, O., Pryor, S., Palmgren, F. and Olsen, E., 2001, *Human exposure to outdoor air pollution*, Pure and Applied Chemistry, Vol. 73 (6), pp. 933-958

Hider, Z. E., Hibberd, S. and Baker, C. J., 1997, *Modelling particulate dispersion in the wake of a vehicle*, Journal of Wind Engineering & Industrial Aerodynamics, Vol. 67&68, pp. 733-744

Hitchins, J., Morawska, L., Wolff, R. C., and Gilbert, D., 2000, *Concentration of submicrometer particles from vehicle emissions near a major road*, Atmospheric Environment, Vol. 34, pp. 51-59

Johnson, L. and Ferreira, L., 2001, *Modelling particle emissions from traffic flows at a freeway in Brisbane, Australia*, Transportation Research Part D, Vol. 6, pp. 357-369

Joumard R., Jost P., Hickman J. and Hassel D., 1995, *Hot passenger car emissions modelling as a function of instantaneous speed and acceleration*, Science of the Total Environment, Vol. 169, pp. 167-174

Junker, M., Kasper, M., Röösli, M., Camenzind, M., Künzli, N., Monn, Ch., Theis, G. and Braun-Fahrländer, Ch., 2000, *Airborne particle number profiles, particle mass distributions and particle-bound PAH concentrations within the city environment of Basel: an assessment as part of the BRISKA Project*, Atmospheric Environment, Vol. 34, pp. 3171-3181

Keady, P.B., 2000, *Getting data you need with particle measurement*, Health and Safety Application Note ITI-075, TSI Inc., Minnesota, USA, http://www.tsi.com/shared/ieg/iti_075.pdf (Reprint from Indoor Environmental Connections, Nov. 2000, Volume 2(1))

Keywood, M. D., Ayers, G. P., Gras, J. L., Gillett, R. W. and Cohen, D. D., 1999, *Relationships between size segregated mass concentration data and ultrafine particle number concentrations in urban areas of Australia*, Atmospheric Environment, Vol. 33, pp. 2907-2913

Kittelson, D. B., 1998, *Engines and nanoparticles: A review*, Journal of Aerosol Science, Vol. 29, pp. 575-588

Kittelson, D. B., Johnson, J., Watts, W., Wei, Q., Drayton, M., Paulsen, D. and Bukowiecki, N., 2000, *Diesel aerosol sampling in the atmosphere*, SAE Technical Paper Series, SAE, Washington D.C., USA

Kukkonen, J., Öttl, D., Almbauer, R. A., Sturm, P. J., Pohjola, M. and Härkönen, J., 2001, *Diagnostic evaluation of two atmospheric dispersion models against a roadside dataset*, contribution to the SATURN project, http:// aix.meng.auth.gr/saturn/annualrep01/, accessed June 2004

Kulmala, M., Vehkamäki, H., Petäjä, T., Dal Maso, M., Lauri, A., Kerminen, V., Birmili, W. and McMurry, P. H., 2004, *Formation and growth rates of ultrafine atmospheric particles: a review of observations*, Journal of Aerosol Science, Vol. 35, pp. 143-176

Longley, I. D., Gallagher, M. W., Dorsey, J. R., Flynn, M., Allan, J. D., Alfarra, M. R. and Inglis, D., 2003, *A case study of aerosol (4.6 nm < $D_p$ < 10 μm) number and mass size distribution measurements in a busy street canyon in Manchester, UK*, Atmospheric Environment, Vol. 37, pp. 1563-1571

Mardaljevic, J. and Rylatt, M., 2003, *Irradiation mapping of complex urban environments: an image-based approach*, Energy and Buildings, Vol. 35, pp. 27-35

Molnár, P., Janhäll, S. and Hallquist, M., 2002, *Roadside measurements of fine and ultrafine particles at a major road north of Gothenburg*, Atmospheric Environment, Vol. 36, pp. 4115–4123

Morawska, L., Vishvakarman, D., Mengersen, K. and Thomas, S., 1999, *Spatial variation of airborne pollutant concentrations in Brisbane, Australia, and its potential impact on population exposure assessment*, Atmospheric Environment, Vol. 36, pp. 3545-3555

Morton, S., 2004, Evidence of the impact of transport on health in *Transport, environment and health in Europe: evidence, initiatives and examples*, World Health Organization, Regional Office for Europe, http://www.who.dk/ eprise/main/who/progs/hcp/UrbanHealthTopics/20020107_1, accessed June 2004

MVA, 1998, *TRIPS/32 Transportation Model*, Tutorial CD, MVA Ltd., Woking, UK

NRC, 1991a, Chapter 2: Framework for assessing exposures to air contaminants, in *Human exposure assessment for airborne pollutants: Advances and opportunities*, pp. 37-52, National Research Council, National Academy of Sciences, Washington, DC

NRC, 1991b, Chapter 5: Survey research methods and exposure assessment, in *Human exposure assessment for airborne pollutants: Advances and opportunities*, pp. 143-168, National Research Council, National Academy of Sciences, Washington, DC

Oglesby L., Künzli N., Röösli M., Braun-Fahrländer C., Mathys P., Stern W., Jantunen M. and Kousa A., 2000, *Validity of ambient levels of fine particles as surrogate for personal exposure to outdoor air pollution - Results EXPOLS-EAS study*, Basel, Journal of the Air & Waste Management Association, Vol. 50, pp. 1251-1261

Pope III, C. A., 2000, *Epidemiology of fine particulate air pollution and human health: Biologic mechanisms and who's at risk?*, Environmental Health Perspectives, Vol. 108(4), pp. 713-723

QMS, 2002, *EViews 4 User's Guide*, Quantitative Micro Software, LLC, USA

QUARG, 1996, *Airborne particulate matter in the United Kingdom*, Third report of the Quality of Urban Air Review Group, Birmingham, UK

Rhodes, C., Sheldon, L., Whitaker, D., Clayton, A., Fitzgerald, K., Flanagan, J., DiGenova, F., Frazier, C. and Hering, S., 1998, *Measuring concentrations of selected air pollutants inside California vehicles*, Executive summary of final report (ARB Contract No. 95-339), California Air Resources Board

Sarnat, J. A., Demokritou, P. and Koutrakis, P., 2003, *Measurement of fine, coarse and ultrafine particles*, Annali dell'Istituto Super Sanità, Vol. 39 (3), pp. 351-355

Sexton, K. and Ryan, R. B., 1988, Assessment of human exposure to air pollution: Methods, measurements, and models, in *Air pollution, the automobile, and public health*, Watson, A.Y., Bates, R.R. and Kennedy, D.(eds.), pp. 207-238, National Academy Press, Washington, DC

StatSoft, 2004, *Electronic Statistics Textbook*, StatSoft Inc., 1984-2004, http:// www.statsoft.com/textbook/stathome.html, accessed June 2004

TSI, 2000, *P-TRAK™ Ultrafine Particle Counter, Theory of Operation*, Application Note ITI-071, TSI Inc., Minnesota, USA

TSI, 2004, *P-TRAK™ Ultrafine Particle Counter (Model 8525), Operation and Service Manual*, TSI Inc., Minnesota, USA

Tuch, T., Wehner, B., Pitz, M., Cyrys, J., Heinrich, J., Kreyling, W. G., Wichmann H.E. and Wiedensohler, A., 2003, *Long-term measurements of size-segregated ambient aerosol in two German cities located 100 km apart*, Atmospheric Environment, Vol. 37, pp. 4687-4700

Turnkey Instruments, 2001, *OSIRIS Environmental Dust Monitor, User manual*, Turnkey Instruments, UK

Turpin K. and Mardaljevic J., 2000, *Saffron Lane safer routes to school study*, Report commissioned by Leicester City Council, Transportation Department

Turpin, K., 2004, *Characterisation of vehicle drive cycles for peak hour traffic: Implications for emissions modelling*, PhD Thesis submitted in June 2004, De Montfort University Leicester, UK

USEPA, 1997, *Environmental Protection Agency, 40 CFR Part 50, National Air Quality Standards for Particulate Matter, Final Rule*, Federal Register, Vol. 62, No. 138, pp. 38651-38760

USEPA, 2001, Chapter 3: Physics and chemistry of particulate matter, in *Particulate matter Volume I*, National Center for Environmental Assessment, US Environmental Protection Agency, http://www.epa.gov/ncea/pdfs/0671v1fm.pdf, accessed June 2004

Wehner, B. and Wiedensohler, A., 2003, *Long term measurements of submicrometer urban aerosols: Statistical analysis for correlations with meteorological conditions and trace gases*, Atmospheric Chemistry and Physics, Vol. 3, pp. 867-879

Wehner, B., Birmili, W., Gnauk, T. and Wiedensohler, A., 2002, *Particle number size distributions in a street canyon and their transformation into the urban background: Measurements and a simple model study*, Atmospheric Environment, Vol. 36, pp. 2215-2223

Weijers, E. P., Even, A., Kos, G. P. A., Groot, A. T. J., Erisman, J. W., and ten Brink, E. M., 2001, *Particulate matter in urban air: Health risks, instrumentation and measurements, and political awareness*, Report (ECN-R-01-002) for research project 7.2745: Stedelijke Luchtkwaliteit, Netherlands Energy Research Foundation, Petten, Netherlands

Weijers, E. P., Khlystov, A. Y., Kos, G. P. A. and Erisman, J. W., 2004, *Variability of particulate matter concentrations along roads and motorways determined by a moving measurement unit*, Atmospheric Environment, Vol. 38, pp. 2993-3002

WHO, 2000, *Air quality guidelines for Europe*, World Health Organization, Regional Office for Europe, Copenhagen, Denmark

Willmott, C. J., 1982, *Some comments on the evaluation of model performance*, Bulletin of the American Meteorological Society, Vol. 63, No. 11, pp. 1309-1313

Wilson, W. E., Mage, D. T., Grant, L. D., 2000, *Estimating separately personal exposure to ambient and nonambient particulate matter for epidemiology and risk assessment: why and how*, Journal of the Air & Waste Management Association, Vol. 50, pp. 1167-1183

Wilson, W. E., Suh, H. H., 1997, *Fine particles and coarse particles: Concentration relationships relevant to epidemiological studies*, Journal of the Air & Waste Management Association, Vol. 47, pp.1238-1249

# *Glossary*

**CPC** — **Condensation Particle Counter.** Device which optically detects very small particles. Prior to the measurement, the size of a particle is increased by condensing vapour onto it so that it 'grows' to sufficient size to be detected optically.

**EViews** — **Statistical analysis software** (QMS, 2002). Special statistical functions from EViews were used in the data analysis, in addition to custom written IDL routines.

**IDL** — **Interactive data language** (Research Systems, Inc.). An extensive set of IDL routines was custom written to aid data processing and analysis.

**Link** — **Section of road on a data collection route** with (near) constant traffic flow and road layout. A new link starts where the traffic flow (i.e. number of vehicles) changes, i.e. due to cars turning off or joining, or where the road layout changes. Both instances typically occur at crossroads and junctions, which results in road links being often terminated by traffic lights or crossroads.

**Macro** — **Custom written VBA (Visual Basic for Applications) routine**.

**MLR or MR** — **Multiple linear regression analysis**, used to identify the main determinants of driver exposure (Chapter 5) and for modelling link and journey exposure (Chapter 6).

**OPC** — **Optical Particle Counter**. Device that detects particles using light scatter. Particles are counted and mass concentrations are estimated using scaling factors.

**OSIRIS** **Environmental dust monitor** (Turnkey Instruments, Ltd.), an Optical particle counter used in the field study to measure mass concentrations of $PM_{2.5}$ and $PM_{10}$ particles. Particle detection range: 0.4 μm - 20 μm.

**PM** **Particulate Matter**. An abbreviation occasionally used in this thesis to collectively refer to particles of all size ranges.

**$PM_{10}$** **Mass concentration** of particles which pass through a size selective inlet with a 50 % efficiency cut-off at **10** μm aerodynamic diameter [$\mu g\, m^{-3}$]. Current air quality standards in the UK are based on this particle size fraction. Particle mass equivalent values for particles < 10 μm, measured during the field study with the OSIRIS monitor, are referred to as $PM_{10}$ data throughout this thesis (see also discussion in Section 3.1.2).

**$PM_{2.5}$** **Mass concentration** of particles which pass through a size selective inlet with a 50 % efficiency cut-off at **2.5** μm aerodynamic diameter [$\mu g\, m^{-3}$]. Particle mass equivalent values for particles < 2.5 μm, measured during the field study with the OSIRIS, are referred to as $PM_{2.5}$ data throughout this thesis (see discussion in Section 3.1.2).

**P-Trak** **Condensation particle counter** (model 8525, TSI Inc.), used during the field study to measure number concentrations of ultrafine particles. Smallest particle size detected: 0.02 μm.

**Run** **Two iterations of a data collection circuit** (route) during the field study. Typically, one run was completed on each route during a data collection session.

**TEOM** **Tapered Element Oscillating Microbalance**. Device widely used in the UK for monitoring mass concentrations of particulate matter. The TEOM measurements are based on the principle of deposited particle mass changing the frequency of mechanical oscillation of a tapered glass tube.

**UF** **Ultrafine Particles.** Particles with an aerodynamic diameter $D_p$ < 0.1 μm. These particles are typically measured as number concentrations [$cm^{-3}$]. Particle concentrations measured during the field study with the P-Trak are referred to as UF or ultrafine particle data throughout this thesis (see discussion in Section 3.1.2).

**VBA** **Visual Basic for Applications** (Microsoft Corporation). Programmable software which aids automation of functions in Microsoft® Excel. A set of programming routines (macros) were written to automate the processing of data during the field study.

Appendix

# A Auxiliary Material for Chapter 3

## A.1 Calibration of speed sensor

The current (mA) output from the speed sensor arrangement was converted into speed data using a calibration function, which was derived from dynamometer tests. Two tests were performed: (1) driving the vehicle against the load supplied by the dynamometer; and, (2) driving it with this load. For both conditions, dynamometer speed and current output from the frequency analogue converter were recorded. By using the average of the speed values registered by the dynamometer under both load conditions for equivalent values of current, and plotting them against the current values, a linear regression line ($r^2$=0.9993) was derived as shown in Figure A-1. During the field campaign, this function was used to convert the measured current values into instantaneous speed data. For further details refer to Turpin (2004).

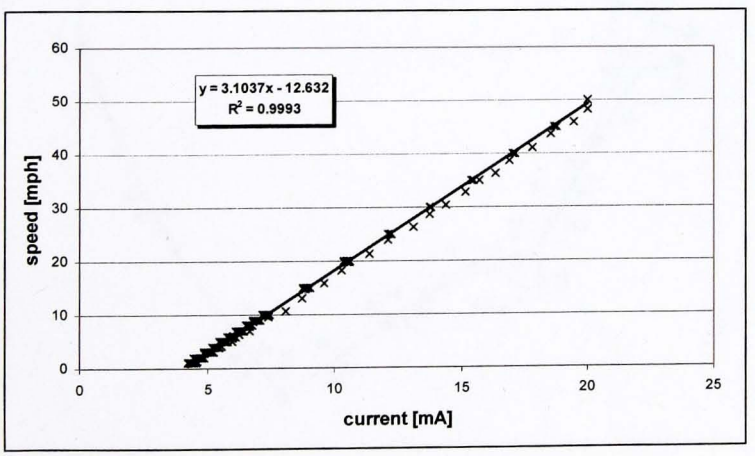


**Figure A-1 Speed calibration function from dynamometer test**

## A.2 Detailed route maps

Figures A-2 to A-4 show maps of the data collection routes. Markers are included to indicate the direction of the traffic flow for each link and the location of traffic lights on the routes. For descriptions of the routes and their locations refer to Section 3.1.3.

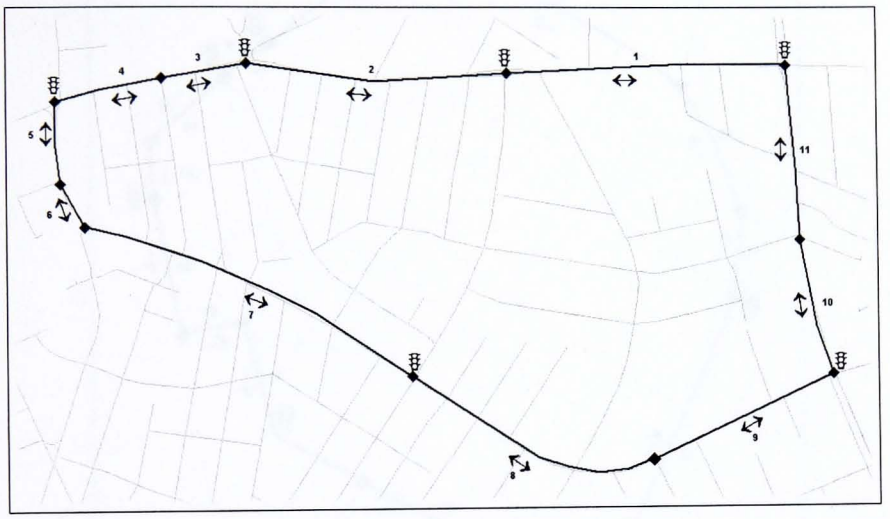


**Figure A-2 Route 1 - Uppingham Road**

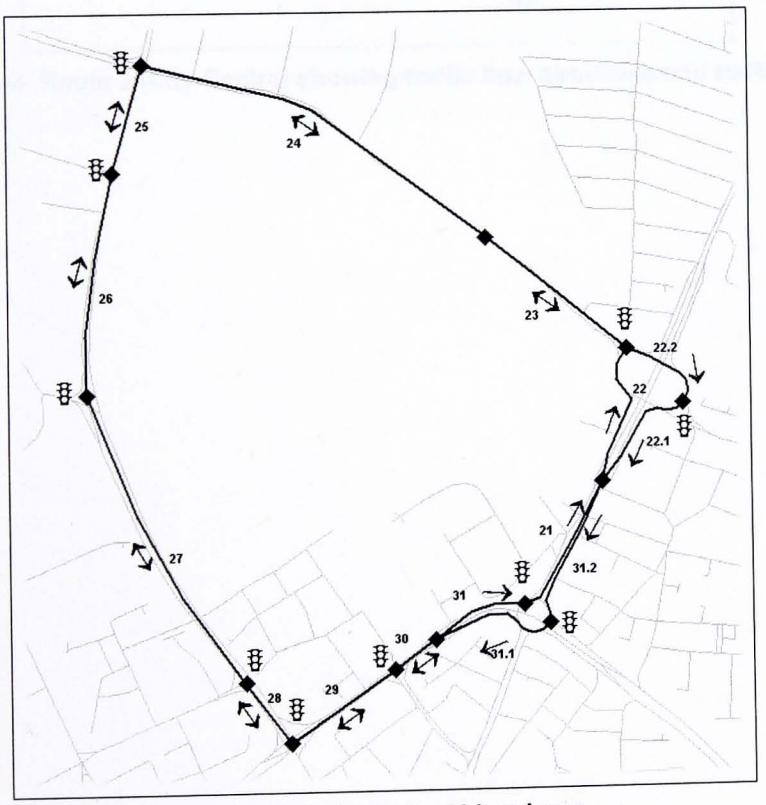


**Figure A-3 Route 2 - Abbey Lane**

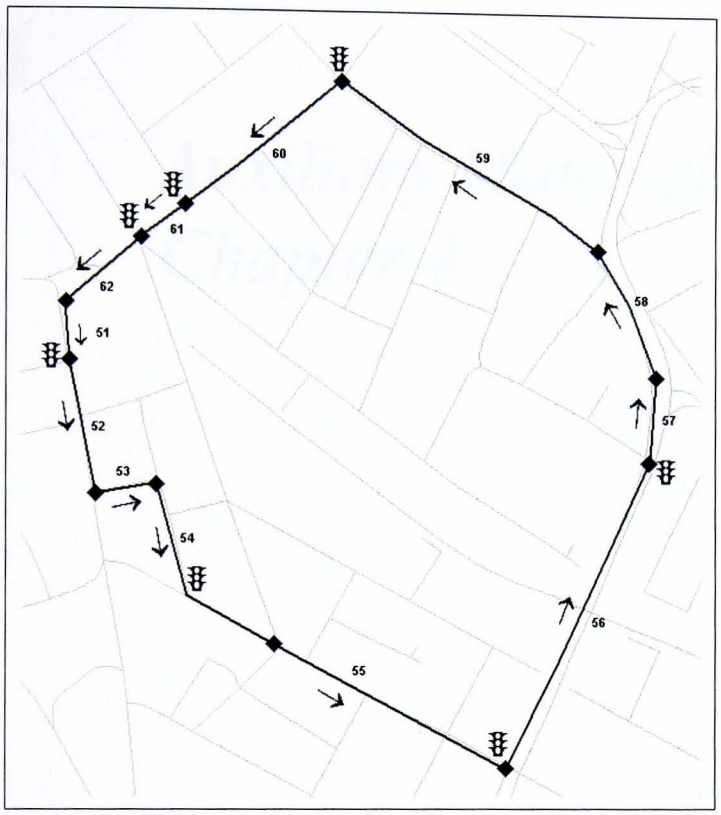


**Figure A-4 Route 3 (City Centre) showing traffic flow directions and traffic lights**

Appendix

# B

# *Auxiliary Material for Chapter 4*

This appendix contains additional data and descriptions of methods for the general data analysis.

# B.1 Descriptive statistics for road links

Table B-1 provides geometric mean and standard deviation values for measured exposure data for the individual road links. For an analysis of the descriptive statistics of the field data set refer to Section 4.2.

| link number | | anti-clockwise | | | | | | clockwise | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | UF [#/cm$^3$] | | PM10 [µg/m$^3$] | | PM2.5 [µg/m$^3$] | | UF [#/cm$^3$] | | PM10 [µg/m$^3$] | | PM2.5 [µg/m$^3$] | |
| | | GM | GSD | GM | GSD | GM | GSD | GM | GSD | GM | GSD | GM | GSD |
| Route 1 Upp. Road | 1 | 54,224 | 1.8 | 35.33 | 2.1 | 13.03 | 1.7 | 76,678 | 1.7 | 47.53 | 2.1 | 17.37 | 2.0 |
| | 2 | 63,821 | 1.8 | 36.89 | 2.0 | 13.04 | 1.7 | 78,424 | 1.7 | 51.90 | 2.1 | 19.03 | 1.9 |
| | 3 | 61,171 | 1.7 | 36.91 | 2.1 | 13.11 | 1.7 | 71,027 | 1.6 | 46.50 | 2.0 | 16.56 | 1.9 |
| | 4 | 68,027 | 1.7 | 40.63 | 2.0 | 14.58 | 1.7 | 65,733 | 1.7 | 41.23 | 2.1 | 15.29 | 2.0 |
| | 5 | 66,789 | 1.7 | 43.35 | 2.1 | 14.86 | 1.7 | 55,375 | 1.7 | 40.00 | 2.1 | 14.73 | 2.0 |
| | 6 | 69,802 | 1.7 | 44.92 | 2.0 | 14.66 | 1.7 | 64,323 | 1.7 | 40.28 | 2.0 | 14.76 | 1.9 |
| | 7 | 47,705 | 1.7 | 33.73 | 2.1 | 11.92 | 1.7 | 57,727 | 1.6 | 37.88 | 2.0 | 14.14 | 1.9 |
| | 8 | 49,453 | 1.8 | 33.69 | 2.0 | 11.93 | 1.7 | 62,254 | 1.7 | 41.08 | 2.0 | 15.36 | 1.8 |
| | 9 | 52,902 | 1.9 | 32.27 | 2.1 | 12.00 | 1.8 | 68,263 | 1.7 | 43.26 | 2.0 | 16.09 | 1.9 |
| | 10 | 44,330 | 1.8 | 31.73 | 2.1 | 11.21 | 1.7 | 70,102 | 1.7 | 43.84 | 2.0 | 15.73 | 1.9 |
| | 11 | 49,100 | 1.8 | 29.84 | 2.1 | 11.36 | 1.7 | 75,768 | 1.7 | 47.34 | 2.1 | 17.45 | 1.9 |
| Route 2 Abbey Lane | 21 | 80,930 | 1.5 | 49.72 | 1.8 | 17.17 | 1.7 | 79,598 | 1.6 | 53.67 | 2.0 | 19.59 | 1.9 |
| | 22 | 80,799 | 1.6 | 50.18 | 1.8 | 16.90 | 1.7 | 82,271 | 1.6 | 55.53 | 2.0 | 20.04 | 1.9 |
| | 23 | 81,131 | 1.6 | 50.65 | 1.9 | 17.02 | 1.7 | 85,485 | 1.7 | 54.71 | 1.9 | 19.48 | 1.8 |
| | 24 | 79,995 | 1.6 | 47.68 | 2.0 | 16.59 | 1.8 | 88,409 | 1.6 | 59.25 | 1.9 | 20.01 | 1.7 |
| | 25 | 81,183 | 1.6 | 51.04 | 1.9 | 16.52 | 1.8 | 88,234 | 1.6 | 60.75 | 1.9 | 21.00 | 1.8 |
| | 26 | 82,413 | 1.6 | 51.79 | 2.0 | 17.17 | 1.7 | 82,440 | 1.6 | 54.53 | 2.0 | 19.22 | 1.7 |
| | 27 | 83,776 | 1.6 | 53.65 | 1.9 | 17.98 | 1.7 | 99,055 | 1.5 | 63.23 | 1.9 | 21.86 | 1.7 |
| | 28 | 77,002 | 1.6 | 47.83 | 1.9 | 17.22 | 1.6 | 98,647 | 1.5 | 63.38 | 1.8 | 22.45 | 1.6 |
| | 29 | 78,535 | 1.6 | 49.60 | 1.9 | 17.38 | 1.7 | 92,513 | 1.6 | 61.60 | 1.9 | 21.87 | 1.7 |
| | 30 | 80,689 | 1.5 | 56.04 | 1.8 | 17.53 | 1.5 | 79,841 | 1.6 | 52.70 | 1.9 | 19.08 | 1.7 |
| | 31 | 79,632 | 1.6 | 50.76 | 1.8 | 17.04 | 1.7 | 84,878 | 1.6 | 55.42 | 1.9 | 19.98 | 1.8 |
| | 32 | - | - | - | - | - | - | 79,279 | 1.6 | 51.60 | 1.9 | 19.01 | 1.8 |
| Route 3 City Centre | 51 | 60,889 | 1.8 | 39.60 | 2.0 | 15.42 | 1.9 | - | - | - | - | - | - |
| | 52 | 60,169 | 1.8 | 40.00 | 2.0 | 15.64 | 1.9 | - | - | - | - | - | - |
| | 53 | 58,700 | 1.8 | 43.55 | 2.0 | 15.83 | 1.9 | - | - | - | - | - | - |
| | 54 | 56,700 | 1.7 | 40.39 | 2.0 | 14.85 | 1.9 | - | - | - | - | - | - |
| | 55 | 61,312 | 1.8 | 38.17 | 2.0 | 14.61 | 1.9 | - | - | - | - | - | - |
| | 56 | 64,439 | 1.9 | 42.09 | 2.0 | 15.34 | 2.0 | - | - | - | - | - | - |
| | 57 | 61,566 | 1.7 | 40.66 | 2.1 | 15.07 | 2.0 | - | - | - | - | - | - |
| | 58 | 70,614 | 1.7 | 50.66 | 2.0 | 17.29 | 1.9 | - | - | - | - | - | - |
| | 59 | 64,929 | 1.7 | 46.23 | 2.0 | 16.13 | 1.9 | - | - | - | - | - | - |
| | 60 | 64,735 | 1.7 | 41.87 | 2.0 | 15.86 | 1.9 | - | - | - | - | - | |
| | 61 | 65,222 | 1.8 | 38.97 | 2.0 | 15.53 | 1.9 | - | - | - | - | - | |
| | 62 | 66,470 | 1.7 | 41.54 | 1.9 | 16.32 | 1.8 | - | - | - | - | - | |

**Table B-1 Geometric mean and standard deviation for all concentration data measured on the individual links**

## B.2 Cross-correlation function for time-lag analysis

The C_CORRELATE function in IDL computes the cross-correlation $P_{xy}(L)$ of two sample populations X and Y as a function of the lag L

$$P_{xy}L = \left\{ \begin{array}{ll} \dfrac{\sum_{k=0}^{N-|L|-1} x_{k+|L|} - \bar{x}y_k - \bar{y}}{\sqrt{\sum_{k=0}^{N-1} x_k - \bar{x}^2 \sum_{k=0}^{N-1} y_k - \bar{y}^2}} & \text{For } L < 0 \\ \dfrac{\sum_{k=0}^{N-L-1} x_k - \bar{x}y_{k+L} - \bar{y}}{\sqrt{\sum_{k=0}^{N-1} x_k - \bar{x}^2 \sum_{k=0}^{N-1} y_k - \bar{y}^2}} & \text{For } L \geq 0 \end{array} \right\} \quad \textbf{(B-1)}$$

where x and y are the means of the sample populations $x = (x_0, x_1, x_2, \ldots, x_{N-1})$ and $y = (y_0, y_1, y_2, \ldots, y_{N-1})$, respectively. This function was used for the investigation of time lagged correlations in the field data in (Section 4.4).

## B.3 Modifying time series data for the correlation analysis

The modifications described in this section were used in the investigation of correlations between time series in Section 4.5.

The short averaging time of the field study measurements, together with the extreme short-term variability in the $PM_{10}$ and $PM_{2.5}$ data, was thought to mask underlying correlations between time-series data for the different size ranges.

In order to investigate whether this is the case, correlation coefficients were recalculated for the data after it had been modified using a combination of two methods.

(1) The time series data was averaged over 6 different time intervals, from 5 to 30 seconds, simulating longer-term measurements.

(2) Using fast Fourier transformation, all periodicities below a certain cut-off value (e.g. 5 seconds) were removed from the time series data, leading to a smoother data line. For further details regarding this methods see Section 4.7. Twelve smoothing settings were used, from 5 to 60. Since this method was applied to each set of time series after it had been modified for the new averaging time, the settings should be considered as frequency factors. For example, for a smoothing setting of 5 means that all data with periodicity above 1/5 Hz is removed, whether it is applied to unmodified field data or a time series that has already been modified to simulate an averaging time of 10 seconds (i.e. only one data value for each 10 second interval).

Colour coded matrices were generated which allowed the visual inspection of the distribution of correlation coefficients as a function of smoothing value and averaging time, as shown in Figure B-1. The colour of the bins in Figures B-1a-c represents the average of the correlation coefficients observed for the settings as given by the associated values on the two axes. The colour of the bins in Figures B-1d-f shows the standard deviation of all correlation coefficients found for the bin. The bins for $x = 0$ and $y = 0$ in each matrix represent the results for the unmodified time-series data.

Although stronger correlations are generally found for the modified time-series data, they are associated with high standard deviation values for ultra-fine particles vs. $PM_{2.5}$ and $PM_{10}$. Only for $PM_{2.5}$ vs. $PM_{10}$ can high correlations with low variability

observed. These are particularly pronounced for long averaging times and smoothing values.

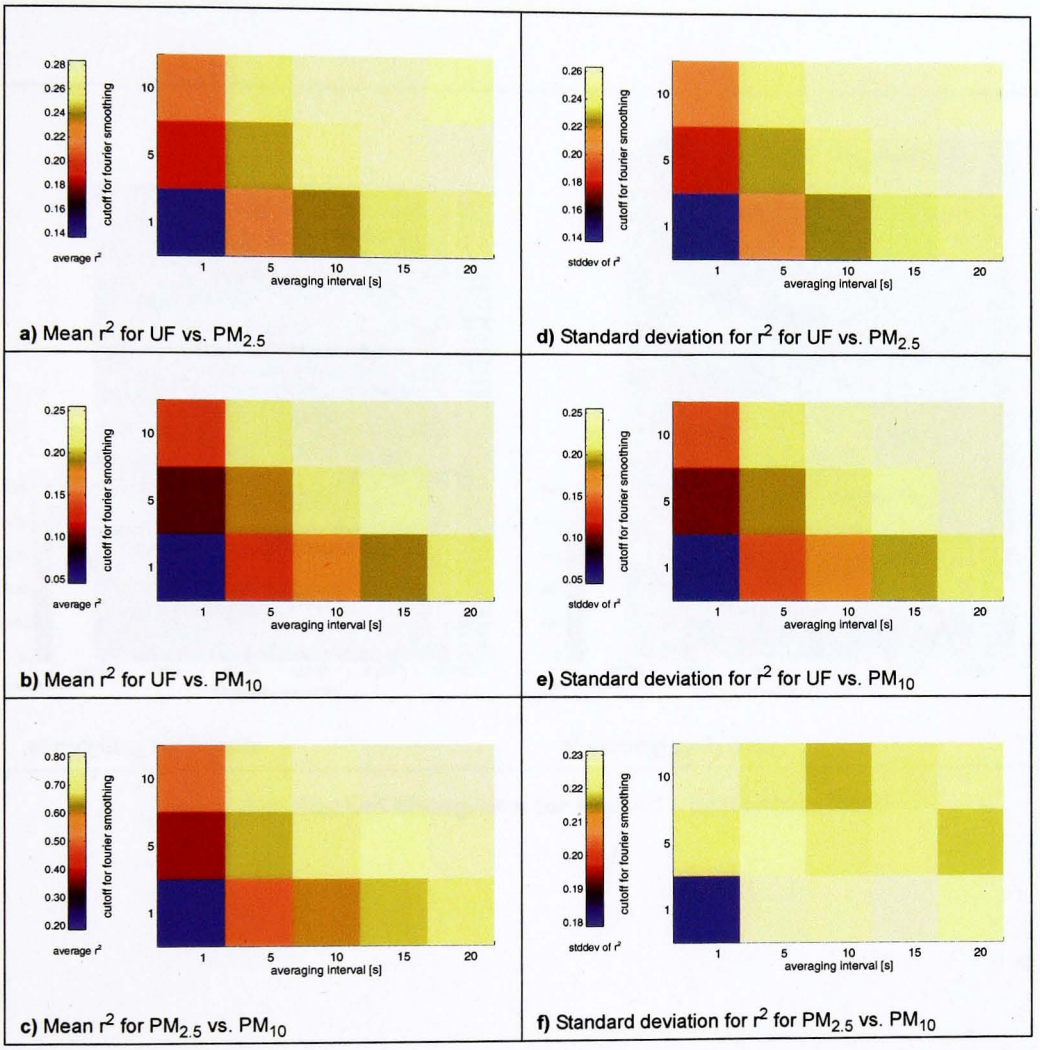


**Figure B-1 Coefficients of correlation between size ranges for various Fourier cut-off values and averaging intervals**

## B.4 Stacked histogram plots for all routes

Figures B-2 to B-4 show stacked histograms for data collected on the field study routes during morning and evening rush hour. For a discussion of the figures refer to Section 4.6.

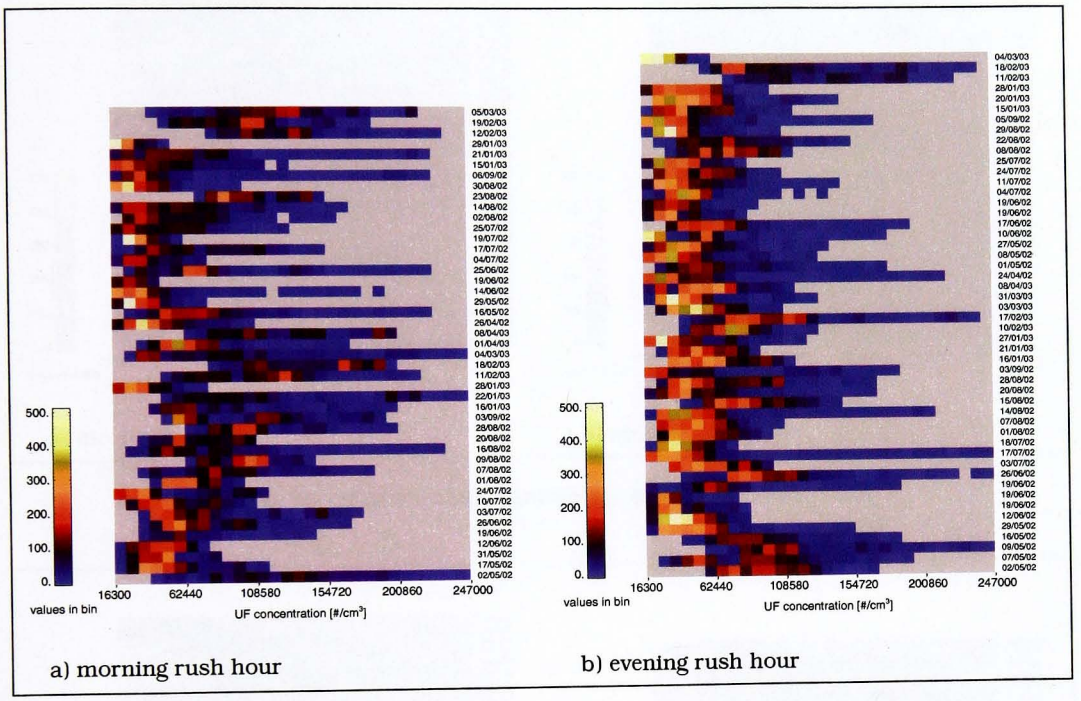


**Figure B-2 Stacked histograms for Route 1 - Uppingham Road**

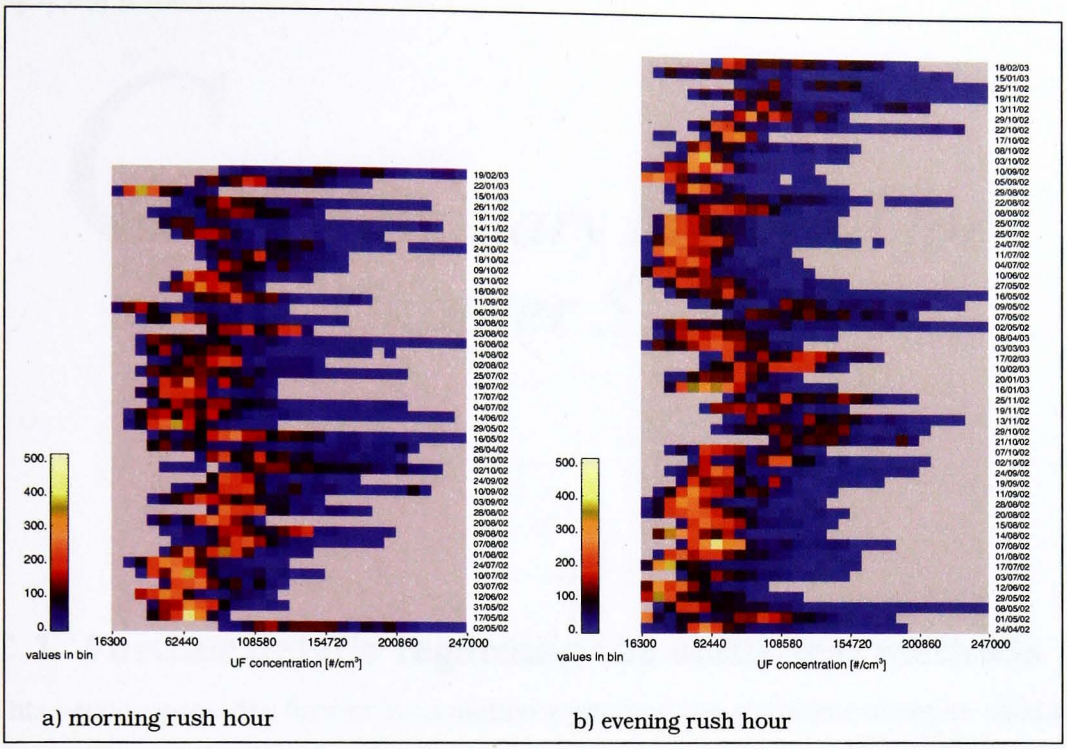


**Figure B-3 Stacked histograms for Route 2 - Abbey Lane**

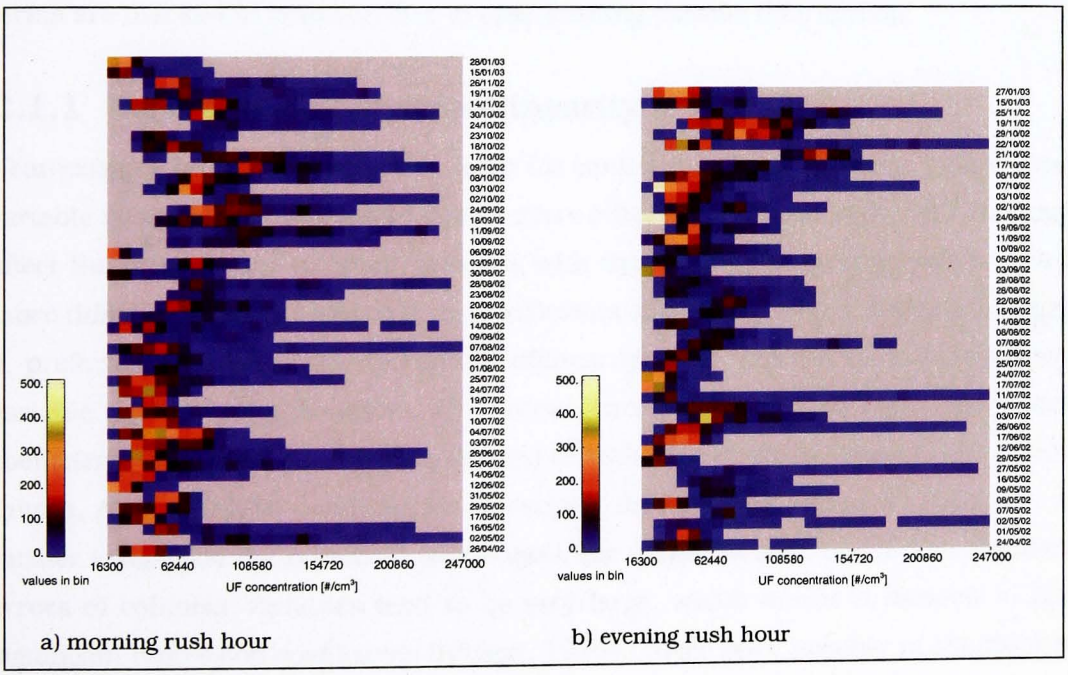


**Figure B-4 Stacked histograms for Route 3 - City Centre**

Appendix

# C Auxiliary Material for Chapter 5

## C.1 Further details regarding the statistical methods

This section provides further information regarding the statistical methods used to identify the main determinants of driver exposure, in addition to the overview of the theoretical background of the statistical methods given in Section 5.2. Important terms are marked in bold typeface to ease locating specific information.

### C.1.1 Near-extreme multicollinearity

Near-extreme multicollinearity can occur for continuous variables if one independent variable is highly correlated with one or more other independent variables. This may affect the relationship of other variables with the dependent variable and makes it more difficult to reliably estimate the coefficients of those variables that are collinear. A preferred way of diagnosing multicollinearity is to regress each independent variable against all the others. The occurrence of a large multiple correlation coefficient (or coefficient of determination) $R^2$ indicates multicollinearity. As a rule of thumb, Allison (1999) recommends to exclude the variables where $R^2 > 0.8$ and to further investigate the relationships between the variables if $R^2 > 0.6$. The standard errors of collinear variables tend to be very large, which makes it difficult to find significant regression coefficients (Allison, 1999). There are a number of solutions to this problem, none of which is perfect since no technique can adequately compensate for the lack of information needed to differentiate between the effects of collinear

variables. Most commonly, one or more collinear variables are removed from the model, or collinear variables are combined into one index variable. When one of two collinear variables is to be removed, it is usually the variable less strongly correlated to the dependent variable.

## C.1.2 Least squares method

The regression coefficients $b$ (in Eq 5-3) are typically calculated using the least squares method, where the squares of the residuals (deviation of modelled values from the regression line) are calculated and the coefficients chosen so that this deviation is smallest, i.e. residuals as close to zero as possible. It is generally acknowledged that the least squares method may not be ideal for non-experimental data sets since it is based on the assumption that deviations between the observations and the fitted line all occur in the y direction, with the values for the independent variables being known precisely (e.g. Everitt and Dunn, 2001). However, non-experimental air pollution data, for example, consists of measured values for both the dependent and independent variables, which means that random variance in the x and y direction should be taken into account. Ayers et al. (2001) point out that if the ordinary least squares method is applied to such data sets, the estimated regression coefficients will be underestimated and the intercept will be overestimated.

## C.1.3 Assessing normality and homoskedasticity

**Probability plots** can be generated where ordered observations are plotted against the appropriate values of the assumed cumulative distribution function (e.g. normal probability plot to check for normality). If the data has the assumed distribution, the plot will show a straight line. For multivariate data, each $p$-dimensional observation $x_i$ can be converted into a **generalized distance**, which, plotted against a chi-squared distribution, will show a straight line if the data set has a multivariate normal distribution (Everitt and Dunn, 2001).

The **Jarque-Bera Statistic** (QMS, 2002) is used to test whether a data series is normally distributed. For this study the statistic was calculated using EViews, as follows

$$JB = \frac{N-k}{6}\left(S^2 + \frac{1}{4}(K-3)^2\right) \qquad \text{(C-1)}$$

where S is the skewness, $K$ the kurtosis and $k$ the number of coefficients used to create the series. The EViews program tests the null hypothesis of 'normal distribution' by comparing the $JB$ statistic to $\chi^2$ distribution with 2 degrees of freedom. A small probability value, as reported by EViews, leads to the rejection of the null hypothesis, i.e the series is not normally distributed (QMS, 2002).

**White's Heteroskedasticity Test** (QMS, 2002) is used to test for homoskedasticity (equal variances) in the residuals from a least squares regression (Eq C-2) by using an auxiliary regression (Eq C-3) as shown here.

$$y_t = b_1 + b_2 x_t + b_3 z_t + e_t \tag{C-2}$$

$$e_t^2 = \alpha_0 + \alpha_1 x_1 + \alpha_2 z_t + \alpha_3 x_t^2 + \alpha_4 z_t^2 + \alpha_5 x_t z_t + v_t \tag{C-3}$$

EViews reports a probability value for $N^*R^2$ checked against a $\chi^2$ distribution. A small probability value leads to the rejection of the null hypothesis, i.e. heteroskedasticity (un-equal variances) is likely.

## C.1.4 Assessing the accuracy of the regression coefficients

**Confidence intervals** provide a range of possible values for each of the coefficients. Even though it is not guaranteed that the true value of the coefficient falls within the calculated range, it can be used with reasonable confidence. Typically, 95% confidence intervals are calculated. However, research results using regression modelling tend to focus more on the hypothesis testing than confidence intervals when assessing regression coefficients (Allison, 1999).

**Hypothesis test** can be used to test (1) whether the derived coefficients are significantly different from zero (**t-statistic**) and (2) whether the predictor variables contribute significantly to the model (**F-statistic**).

**(1) t-statistic**: Regression coefficients for the individual variables can be very small (though they are rarely exactly zero), which could be due to random errors generating artifactual relationships when the true coefficient is zero. To test whether this is the case, a t-statistic is calculated for the coefficient by dividing it by its standard error. The associated p-value gives the probability of estimating a coefficient that much different from zero if a true relationship does not exist, i.e. if p is very small, it is

unlikely that the true coefficient is zero. Typical limit values used are $p < 0.05$ ('coefficient significantly different from zero') and $p < 0.01$ ('coefficient highly significant').

**(2) F-statistic**: In order to assess whether the included predictor variables contribute significantly to the calculated regression model, an F-statistic is calculated for the exclusion of each independent variable $x_p$ as follows

$$F = \frac{(RSS_{\text{excl\_xp}} - RSS_{all})/df_1}{RSS_{all}/df_2} \qquad \text{(C-4)}$$

where $RSS_{excl\_xp}$ is the residual sum of squares for the model excluding variable $x_p$ and $RSS_{all}$ is the residual sum of squares of the original model. For the exclusion of one variable, the degrees of freedom are $df_1 = 1$ and $df_2 = n - (p + 1)$. Based on the F-distribution, these values can be used to derive a p-value which indicates the probability that the excluded variable does not contribute significantly to the model, over and above the other independent variables. The smaller the p-value the more likely it is that the excluded variable is significant for the original model. Typically used p-values are 0.05 ('significant') and 0.01 ('highly significant') (Everitt and Dunn, 2001). In the case of categorical variables, i.e. where dummy variables are coded for the various categories, hypothesis tests should be carried out jointly for the exclusion of all dummy columns for that variable (Allison, 1999).

Additionally, an $\mathbf{R^2}$ can be calculated for each variable, which indicates to what extent the variable explains the variability in the dependent variable. This $R^2$ is equivalent to the ratio of the between groups sum of squares (*RSS* for only one variable included) to the total sum of squares.

## C.2 Considerations for the conversion of the circular variable 'wind direction'

Since wind direction is a circular variable, measured in degrees from 0 to 360, its divergence from zero does not increase continuously and thus requires conversion to be useable as an independent variable in multiple linear regression analysis.

In order to removed the discontinuity of the variable, Chaloulakou et al. (2003) introduce a wind direction index which is calculated based on the average particle concentrations (here $PM_{10}$) measured at the main 16 wind directions. This method assumes that the wind direction function is maximised by the wind direction that corresponds to the highest particle concentration measurements.

This method was considered unsuitable for the multiple regression analysis described in Chapter 5 since it was based on an assumption regarding the relationship between wind direction and particle concentrations, and it was this relationship that was the target of the analysis. It was therefore decided to investigate the available wind direction data in more detail in order to devise a conversion method appropriate for this analysis.

A windrose was generated as a first visualisation of the wind speed and wind direction data available for analysis. The wind rose is shown in Figure C-1 together with a histogram of the wind direction values.

The histogram shows that wind directions around 200° occurred most frequently with values decreasing with increasing distance from this angle. Setting the most often recorded wind direction to 0 (here 199°) and calculating the distance in degrees of all other values from this value, lead to a new 'normalized' wind direction variable (see scatterplot in Figure C-2). This variable was slightly negatively correlated to wind speed ($r^2 = 0.03$). In fact, when the correlation between wind speed and wind direction was investigated for all 'normalized' wind direction variables, i.e. by converting wind direction in the described way for each of the 360 angles and calculating the correlation coefficient r between wind speed and the converted variable, the strongest correlations were found for angles between 170° and 200° (see Figure C-2). Although the observed correlation factors are small ($|r_{max}| = 0.188$), the fact that the strongest (negative) correlations are observed when the most frequently recorded wind direction angles are used for conversion (i.e. set to 0) indicates that

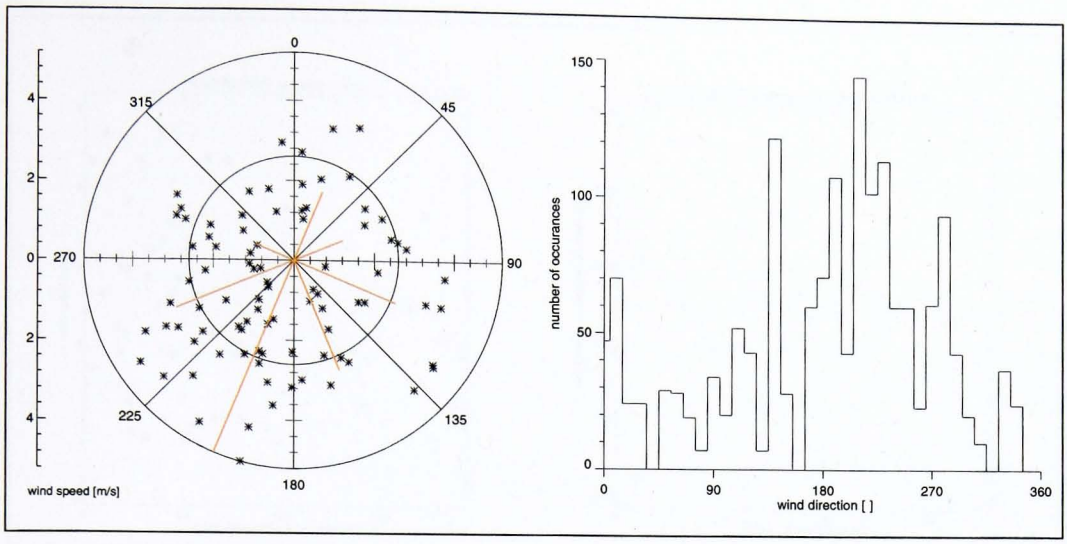


**Figure C-1 Wind rose and histogram for wind direction variable**

*Note, the orange line indicates the proportion of measurements falling within the each of the eight dummy categories coded for the categorical variable wind direction.*

winds coming from the most often recorded direction tend to be slightly stronger. This finding was itself useful for the interpretation of the analysis results in Chapter 5.

The conversion method produces a wind direction variable (distance from a selected angle) that is continuous rather than circular and can thus be generally used in regression analyses. However, since the inclusion of this converted wind direction variable would have meant to investigate the effect of the angular displacement of the wind direction from the most frequently recorded value on the dependent variable rather than the effect of wind direction in general, it was decided not to use this 'normalized' variable.

Instead, wind direction (WD) was treated as a categorical variable and coded with eight dummy columns, i.e. one category for each 45 degree section of the compass (cat0: NNE, $0° < \alpha \leq 45°$; cat1: ENE, $45° < \alpha \leq 90°$; etc.). It was accepted that this approach leads to a certain loss of detail, i.e. values are included as belonging to relatively wide categories rather than actual measured values. However, it allows wind direction to be included as a truly 'independent' variable so that it's effects on driver exposure could be investigated without having to disaggregate the effect of the most measured values or particle concentrations from the results.

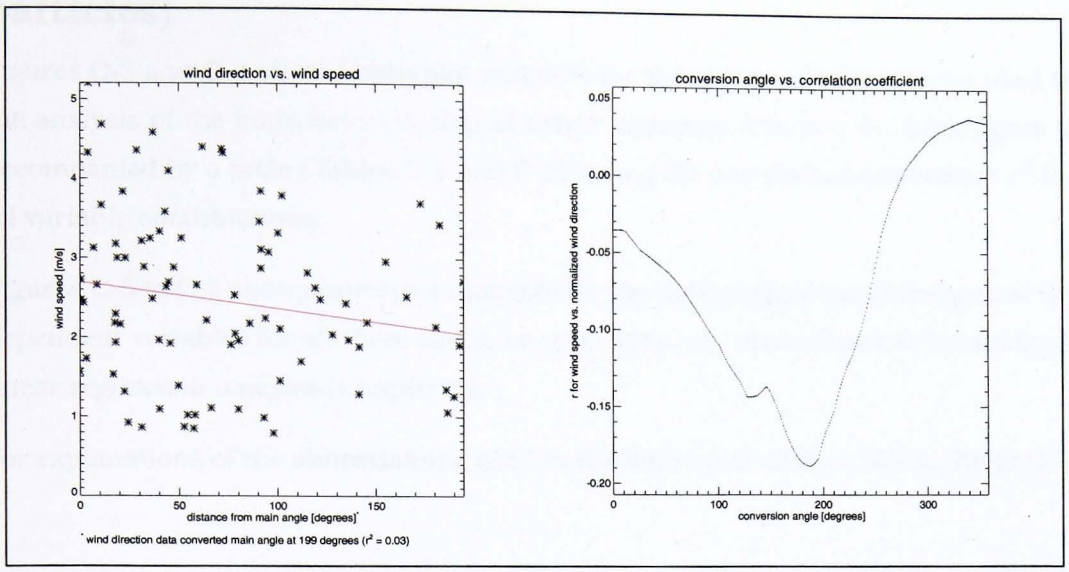


**Figure C-2 Scatterplot for wind speed vs. converted wind direction for the most frequently measured angle (left) and correlation coefficients for wind speed vs. converted wind direction for all angles**

## C.3 Scatterplots and correlation matrices (ultrafine particles)

Figures C-3 and C-4 show scatterplot matrices for the independent variables used in the analysis of the main determinants of driver exposure (Chapter 5). Each figure is accompanied by a table (Tables C-1 and C-2) listing the correlation coefficients $r^2$ for all variable combinations.

Figures C-5 to C-7 show scatterplot matrices for the independent variables against the dependent variables for all three cases, i.e. ATE, ATEK, AC, investigated in the multiple linear regression analysis (Chapter 5).

For explanations of the abbreviations used in the tables and charts refer to Table C-3.

| | LL | tFL | tSP | SPL | RF | | | | C | LED | | | | RSE |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| LL | 1.0000 | -0.1667 | 0.1120 | 0.1951 | -0.1553 | -0.1687 | 0.1772 | 0.1847 | 0.1953 | -0.2573 | 0.1514 | 0.1812 | -0.1077 | 0.1817 |
| tFL | -0.1667 | 1.0000 | -0.0371 | 0.1429 | -0.2663 | 0.2128 | 0.1359 | 0.1037 | 0.0558 | 0.5245 | -0.0560 | 0.0732 | -0.1039 | -0.3506 |
| tSP | 0.1120 | -0.0371 | 1.0000 | 0.3462 | -0.4415 | 0.1914 | 0.2132 | 0.0783 | -0.1416 | 0.0765 | 0.0912 | 0.0259 | 0.0348 | -0.0472 |
| SPL | 0.1951 | 0.1429 | 0.3462 | 1.0000 | -0.1073 | -0.2596 | 0.3249 | 0.2776 | 0.4006 | -0.0787 | 0.2874 | -0.0787 | -0.0332 | -0.1265 |
| RF | -0.1553 | -0.2663 | -0.4415 | -0.1073 | 1.0000 | -0.2824 | -0.1477 | -0.1436 | 0.0177 | -0.0856 | 0.0342 | -0.0856 | -0.0362 | -0.3205 |
| | -0.1687 | 0.2128 | 0.1914 | -0.2596 | -0.2824 | 1.0000 | -0.3573 | -0.3473 | -0.2844 | 0.3030 | -0.3426 | 0.0905 | -0.0875 | -0.4189 |
| | 0.1772 | 0.1359 | 0.2132 | 0.3249 | -0.1477 | -0.3573 | 1.0000 | -0.1816 | 0.3740 | -0.1083 | 0.2904 | -0.1083 | -0.0457 | 0.1461 |
| | 0.1847 | 0.1037 | 0.0783 | 0.2776 | -0.1436 | -0.3473 | -0.1816 | 1.0000 | 0.3373 | -0.1053 | 0.1079 | 0.1870 | -0.0445 | 0.1581 |
| C | 0.1953 | 0.0558 | -0.1416 | 0.4006 | 0.0177 | -0.2844 | 0.3740 | 0.3373 | 1.0000 | -0.1964 | 0.0486 | 0.0186 | -0.0830 | -0.0551 |
| LED | -0.2573 | 0.5245 | 0.0765 | -0.0787 | -0.0856 | 0.3030 | -0.1083 | -0.1053 | -0.1964 | 1.0000 | -0.2737 | -0.0627 | -0.0265 | -0.2349 |
| | 0.1514 | -0.0560 | 0.0912 | 0.2874 | 0.0342 | -0.3426 | 0.2904 | 0.1079 | 0.0486 | -0.2737 | 1.0000 | -0.2737 | -0.1156 | 0.0020 |
| | 0.1812 | 0.0732 | 0.0259 | -0.0787 | -0.0856 | 0.0905 | -0.1083 | 0.1870 | 0.0186 | -0.0627 | -0.2737 | 1.0000 | -0.0265 | 0.1834 |
| | -0.1077 | -0.1039 | 0.0348 | -0.0332 | -0.0362 | -0.0875 | -0.0457 | -0.0445 | -0.0830 | -0.0265 | -0.1156 | -0.0265 | 1.0000 | 0.1128 |
| RSE | 0.1817 | -0.3506 | -0.0472 | -0.1265 | -0.3205 | -0.4189 | 0.1461 | 0.1581 | -0.0551 | -0.2349 | 0.0020 | 0.1834 | 0.1128 | 1.0000 |

**Table C-1 Correlation matrix for the link specific independent variables**

| | S | TOD | PM_BG | T | WS | WD | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| S | 1.0000 | 0.0048 | 0.0352 | 0.6590 | 0.0709 | -0.0371 | -0.2457 | -0.2091 | 0.1082 | 0.1510 | 0.0622 | 0.0574 |
| TOD | 0.0048 | 1.0000 | 0.0391 | 0.3696 | 0.3010 | -0.0422 | -0.0494 | -0.1259 | -0.0820 | 0.1118 | 0.0462 | -0.0755 |
| PM_BG | 0.0352 | 0.0391 | 1.0000 | -0.2194 | -0.1450 | 0.1633 | -0.1674 | -0.0187 | -0.0595 | -0.0603 | -0.0880 | 0.0137 |
| T | 0.6590 | 0.3696 | -0.2194 | 1.0000 | 0.2363 | -0.1205 | -0.1970 | -0.2235 | 0.1420 | 0.1719 | 0.1607 | 0.0289 |
| WS | 0.0709 | 0.3010 | -0.1450 | 0.2363 | 1.0000 | -0.0303 | 0.1306 | -0.1633 | 0.1336 | 0.0828 | -0.1191 | -0.0792 |
| WD | -0.0371 | -0.0422 | 0.1633 | -0.1205 | -0.0303 | 1.0000 | -0.0737 | -0.0818 | -0.1327 | -0.1184 | -0.0967 | -0.0496 |
| | -0.2457 | -0.0494 | -0.1674 | -0.1970 | 0.1306 | -0.0737 | 1.0000 | -0.1121 | -0.1817 | -0.1621 | -0.1325 | -0.0679 |
| | -0.2091 | -0.1259 | -0.0187 | -0.2235 | -0.1633 | -0.0818 | -0.1121 | 1.0000 | -0.2019 | -0.1801 | -0.1472 | -0.0754 |
| | 0.1082 | -0.0820 | -0.0595 | 0.1420 | 0.1336 | -0.1327 | -0.1817 | -0.2019 | 1.0000 | -0.2920 | -0.2387 | -0.1223 |
| | 0.1510 | 0.1118 | -0.0603 | 0.1719 | 0.0828 | -0.1184 | -0.1621 | -0.1801 | -0.2920 | 1.0000 | -0.2129 | -0.1091 |
| | 0.0622 | 0.0462 | -0.0880 | 0.1607 | -0.1191 | -0.0967 | -0.1325 | -0.1472 | -0.2387 | -0.2129 | 1.0000 | -0.0891 |
| | 0.0574 | -0.0755 | 0.0137 | 0.0289 | -0.0792 | -0.0496 | -0.0679 | -0.0754 | -0.1223 | -0.1091 | -0.0891 | 1.0000 |

**Table C-2 Correlation matrix for the independent variables based on meteorological parameters**

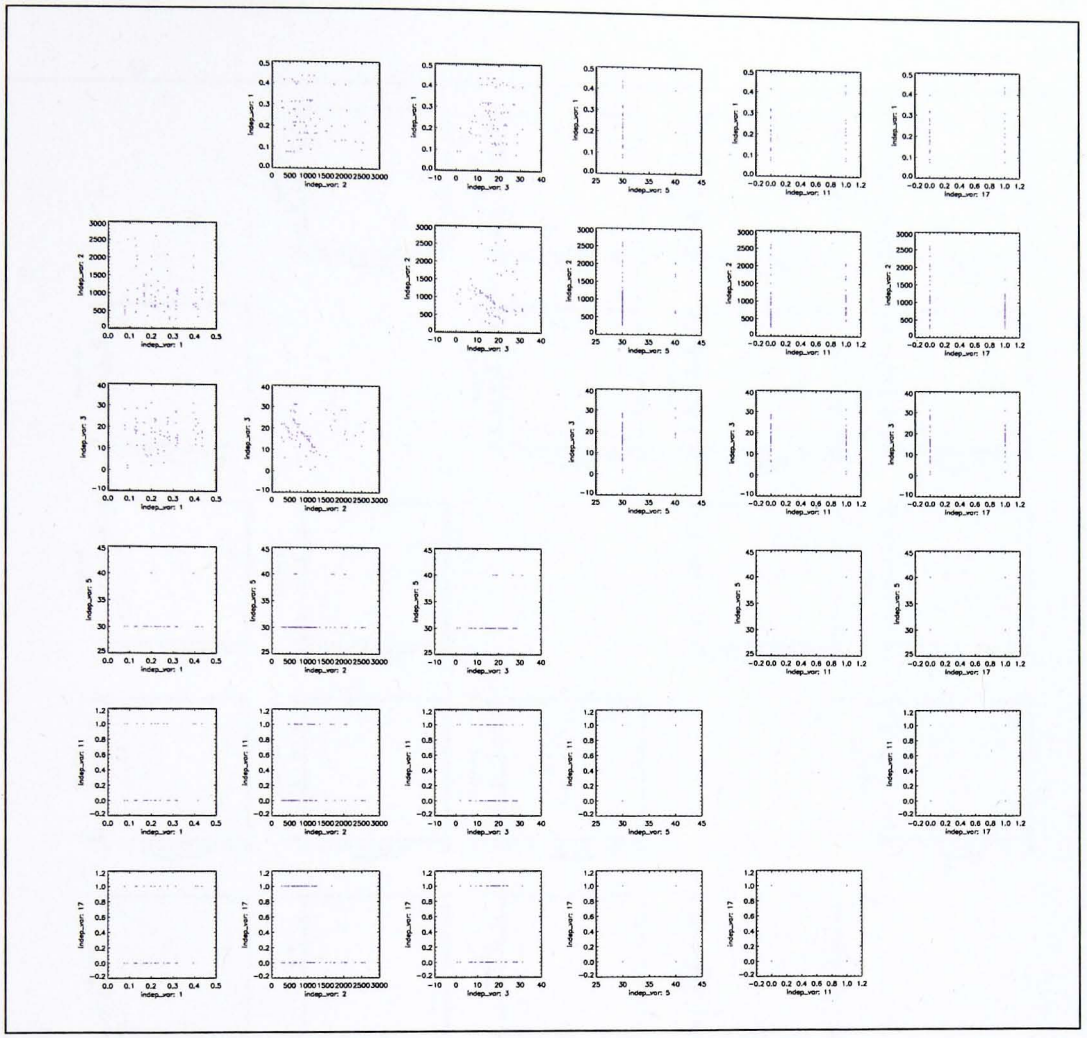

**Figure C-3 Scatterplots for the link specific independent variables**

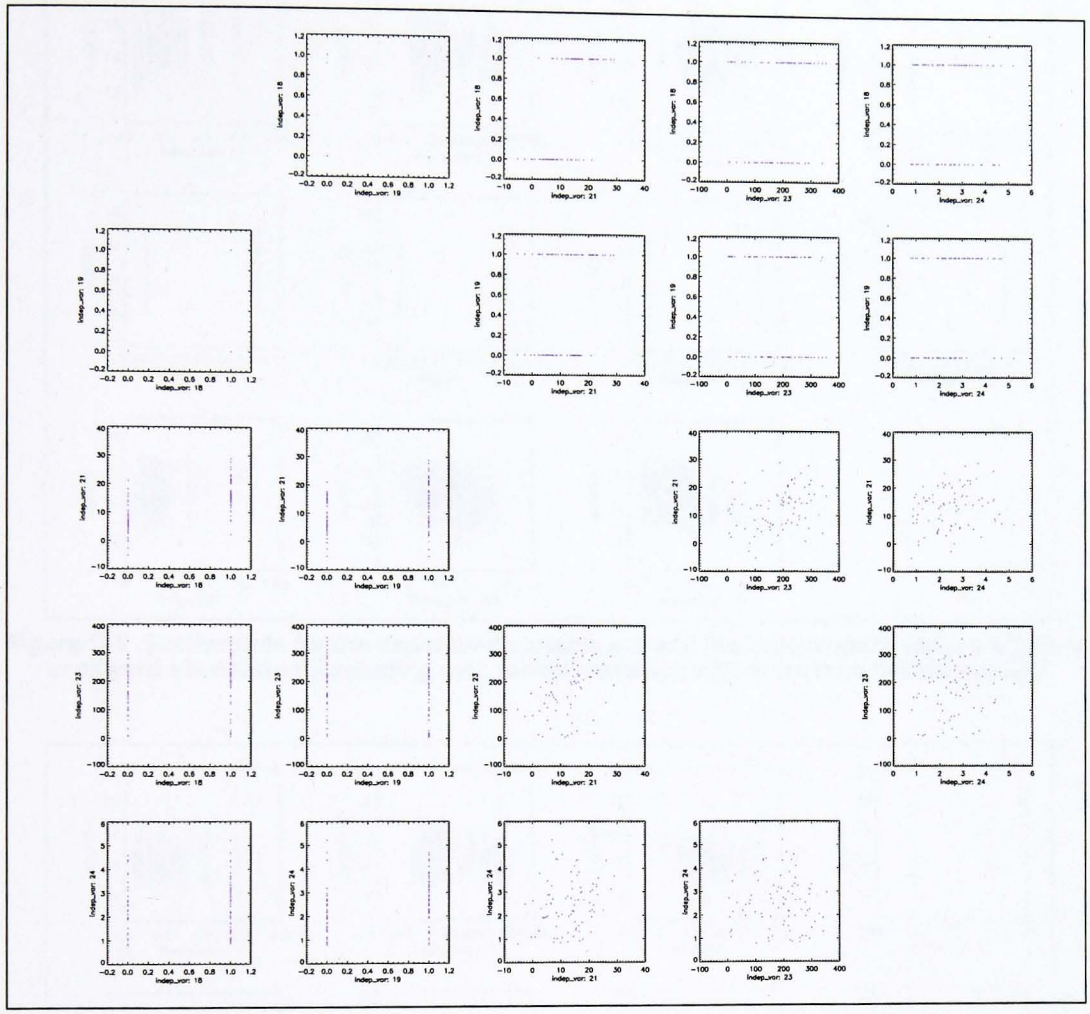

**Figure C-4 Scatterplots for the independent variables based on meteorological parameters**

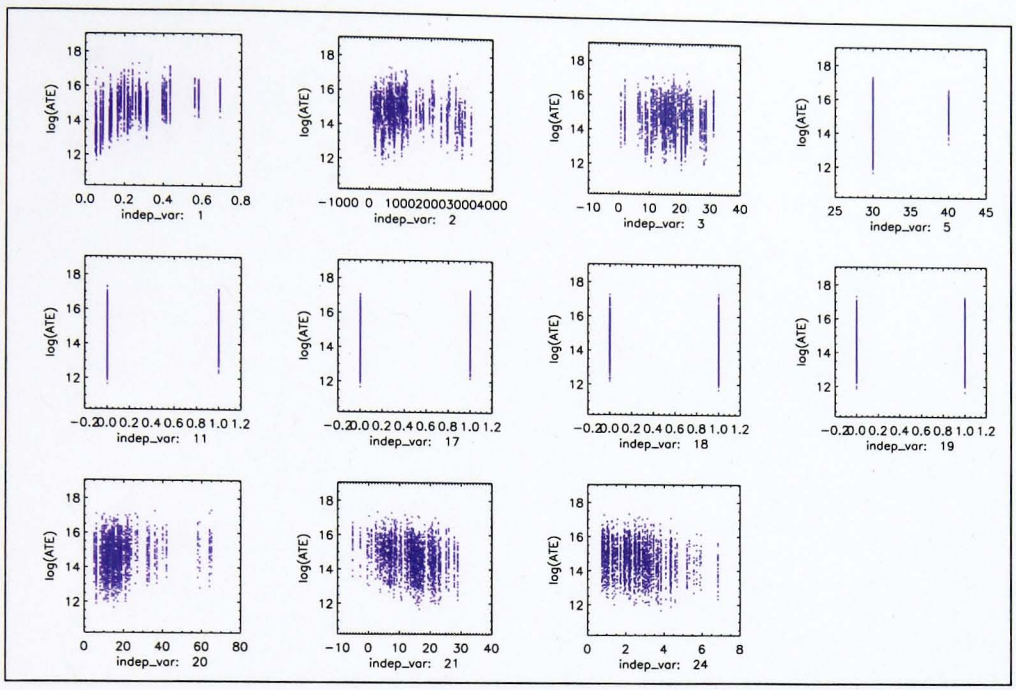


**Figure C-5 Scatterplots for the dependent variable ATE and the independent variables before backward elimination (excluding categorical variables with more than two categories)**

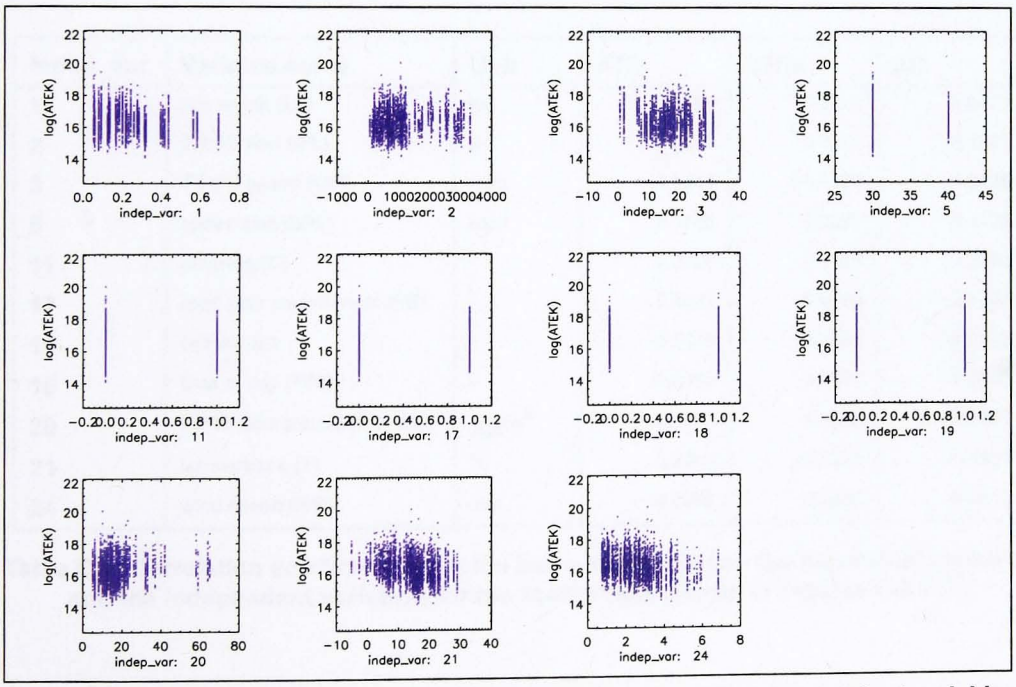


**Figure C-6 Scatterplots for the dependent variable ATEK and the independent variables before backward elimination (excluding categorical variables with more than two categories)**

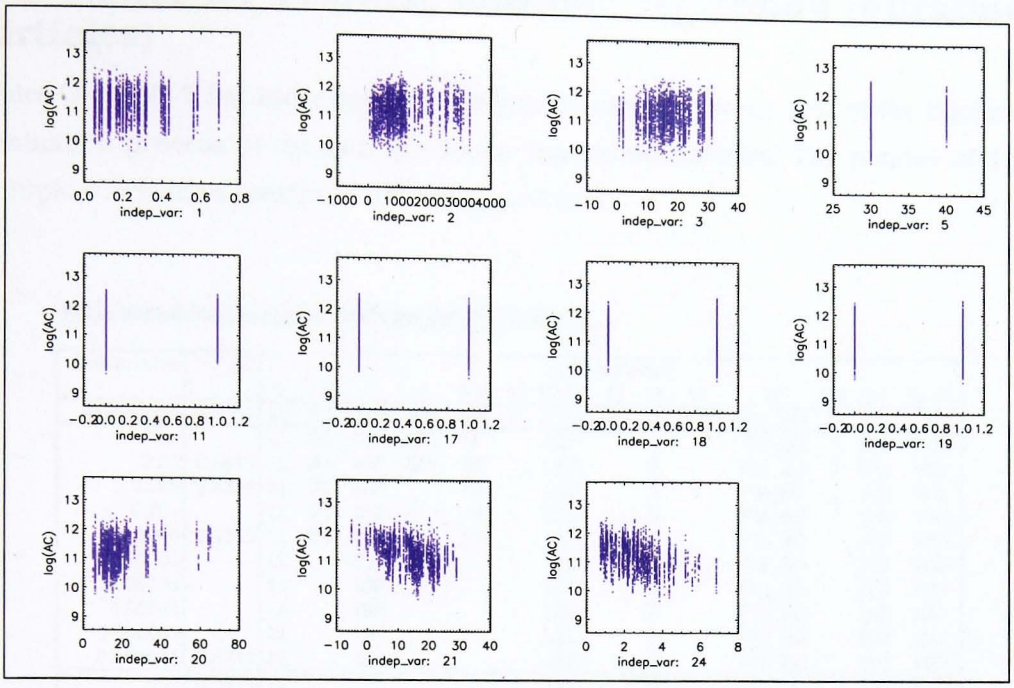


**Figure C-7 Scatterplots for the dependent variable AC and the independent variables before backward elimination (excluding categorical variables with more than two categories)**

| Indep_var | Variable name | Unit | ATE | ATEK | AC |
|---|---|---|---|---|---|
| 1 | link length (LL) | km | 0.3268 | -0.2514 | 0.0073 |
| 2 | TRIPS flow (tFL) | #/h | -0.0013 | 0.0881 | 0.1289 |
| 3 | TRIPS speed (tSP) | mph | -0.0848 | -0.1380 | 0.0439 |
| 5 | speed limit (SPL) | mph | 0.1148 | 0.0027 | 0.1135 |
| 11 | crossing (C) | - | 0.0729 | -0.0057 | 0.1096 |
| 17 | road side environment (RE) | - | 0.0623 | -0.0493 | -0.1394 |
| 18 | season (S) | - | -0.2211 | -0.2341 | -0.3232 |
| 19 | time of day (TOD) | - | -0.0956 | -0.1321 | -0.2450 |
| 20 | PM10 background (PM_BG) | $\mu g/m^3$ | 0.1643 | 0.1538 | 0.2625 |
| 21 | temperature (T) | °C | -0.2805 | -0.2961 | -0.4986 |
| 24 | wind speed (WS) | m/s | -0.2030 | -0.2245 | -0.4013 |

**Table C-3 Correlation coefficients r for the correlation between the dependent variables and the independent variables for the scatterplots shown in Figures C-5 - C-7**

## C.4 Tables for stepwise multiple regression (ultrafine particles)

Tables C-4 to C-7 list independent variables included at every step of the backward elimination process of the multiple linear regression analysis. The $p$-value and the multiple correlation coefficient $R^2$ are also shown.

**backward elimination steps for pt 'ate_all'**

| limit p-value | R2 | vars included | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 5 | 6-10 | 11 | 13-16 | 17 | 18 | 19 | 20 | 21 | 24 | 33-39 |
| all included | 0.4412 | LL | tFL | tSP | SPL | RF | C | LED | RE | S | TOD | PM_BG | T | WS | WD |
| 0.05 | | LL | tFL | tSP | SPL | RF | | LED | | S | | PM_BG | T | WS | WD |
| 0.01 | 0.4411 | LL | tFL | tSP | SPL | RF | | LED | | S | | PM_BG | T | WS | WD |
| 0.005 | 0.4387 | LL | tFL | tSP | | RF | | LED | | S | | PM_BG | T | WS | WD |
| 0.001 | | LL | tFL | tSP | | RF | | LED | | S | | PM_BG | | WS | WD |
| 0.0005 | 0.4319 | LL | tFL | tSP | | RF | | LED | | S | | PM_BG | | WS | WD |
| 0.0001 | | LL | | tSP | | | | LED | | S | | PM_BG | | WS | WD |
| 0.00005 | | LL | | tSP | | | | LED | | S | | PM_BG | | WS | WD |
| 0.00001 | | LL | | tSP | | | | LED | | S | | PM_BG | | WS | WD |
| 0.000005 | | LL | | tSP | | | | LED | | S | | PM_BG | | WS | WD |
| 0.000001 | 0.4241 | LL | | tSP | | | | LED | | S | | PM_BG | | WS | WD |
| end | 0.3914 | LL | | | | | | LED | | S | | | | WS | WD |

**backward elimination steps for pt 'atek_all'**

| limit p-value | R2 | vars included | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 5 | 6-10 | 11 | 13-16 | 17 | 18 | 19 | 20 | 21 | 24 | 33-39 |
| all included | 0.4162 | LL | tFL | tSP | SPL | RF | C | LED | RE | S | TOD | PM_BG | T | WS | WD |
| 0.05 | | LL | tFL | | | RF | C | LED | | S | | PM_BG | T | WS | WD |
| 0.01 | | LL | tFL | | | RF | C | LED | | S | | PM_BG | T | WS | WD |
| 0.005 | 0.4147 | LL | tFL | | | RF | C | LED | | S | | PM_BG | T | WS | WD |
| 0.001 | | LL | tFL | | | RF | C | LED | | S | | PM_BG | | WS | WD |
| 0.0005 | 0.4109 | LL | tFL | | | RF | C | LED | | S | | PM_BG | | WS | WD |
| 0.0001 | | LL | tFL | | | RF | | LED | | S | | PM_BG | | WS | WD |
| 0.00005 | | LL | tFL | | | RF | | LED | | S | | PM_BG | | WS | WD |
| 0.00001 | | LL | tFL | | | RF | | LED | | S | | PM_BG | | WS | WD |
| 0.000005 | 0.4060 | LL | tFL | | | RF | | LED | | S | | PM_BG | | WS | WD |
| 0.000001 | 0.3888 | LL | | | | RF | | LED | | S | | | | WS | WD |
| end | 0.3642 | LL | | | | | | LED | | S | | | | WS | WD |

**backward elimination steps for pt 'ac_all'**

| limit p-value | R2 | vars included | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 5 | 6-10 | 11 | 13-16 | 17 | 18 | 19 | 20 | 21 | 24 | 33-39 |
| all included | 0.4548 | LL | tFL | tSP | SPL | RF | C | LED | RE | S | TOD | PM_BG | T | WS | WD |
| 0.05 | 0.4527 | | | | SPL | RF | | LED | RE | S | TOD | PM_BG | T | WS | WD |
| 0.01 | | | | | | RF | | LED | RE | S | TOD | PM_BG | T | WS | WD |
| 0.005 | 0.4512 | | | | | RF | | LED | RE | S | TOD | PM_BG | T | WS | WD |
| 0.001 | | | | | | RF | | | RE | S | TOD | PM_BG | T | WS | WD |
| 0.0005 | 0.4453 | | | | | RF | | | RE | S | TOD | PM_BG | T | WS | WD |
| 0.0001 | | | | | | RF | | | RE | | | PM_BG | T | WS | WD |
| 0.00005 | 0.4390 | | | | | RF | | | RE | | | PM_BG | T | WS | WD |
| 0.00001 | | | | | | RF | | | | | | PM_BG | T | WS | WD |
| 0.000005 | | | | | | RF | | | | | | PM_BG | T | WS | WD |
| 0.000001 | 0.4328 | | | | | RF | | | | | | PM_BG | T | WS | WD |
| end | 0.3973 | | | | | | | | | | | | T | WS | WD |

**Table C-4 Results for steps of backward elimination analysis for un-grouped ultrafine particle concentration data**

**backward elimination steps for pt 'ate_summer'**

| limit p-value | R2 | vars included | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 5 | 6-10 | 11 | 13-16 | 17 | 18 | 19 | 20 | 21 | 24 | 33-39 |
| all included | 0.4271 | LL | tFL | tSP | SPL | RF | C | LED | RE | | TOD | PM_BG | T | WS | WD |
| 0.05 | 0.4259 | LL | | tSP | SPL | RF | | LED | RE | | TOD | PM_BG | T | WS | WD |
| 0.01 | | LL | | tSP | SPL | RF | | LED | | | | PM_BG | | WS | WD |
| 0.005 | 0.4224 | LL | | tSP | SPL | RF | | LED | | | | PM_BG | | WS | WD |
| 0.001 | | LL | | tSP | | RF | | LED | | | | PM_BG | | WS | |
| 0.0005 | 0.4047 | LL | | tSP | | RF | | LED | | | | PM_BG | | WS | |
| 0.0001 | | LL | | tSP | | | | LED | | | | PM_BG | | WS | |
| 0.00005 | | LL | | tSP | | | | LED | | | | PM_BG | | WS | |
| 0.00001 | | LL | | tSP | | | | LED | | | | PM_BG | | WS | |
| 0.000005 | | LL | | tSP | | | | LED | | | | PM_BG | | WS | |
| 0.000001 | 0.3929 | LL | | tSP | | | | LED | | | | PM_BG | | WS | |
| end | 0.3676 | LL | | | | | | LED | | | | PM_BG | | WS | |

**backward elimination steps for pt 'ate_winter'**

| limit p-value | R2 | vars included | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 5 | 6-10 | 11 | 13-16 | 17 | 18 | 19 | 20 | 21 | 24 | 33-39 |
| all included | 0.4382 | LL | tFL | tSP | SPL | RF | C | LED | RE | | TOD | PM_BG | T | WS | WD |
| 0.05 | 0.4329 | LL | tFL | | | RF | | LED | | | | | T | WS | WD |
| 0.01 | | LL | tFL | | | RF | | LED | | | | | | WS | WD |
| 0.005 | | LL | tFL | | | RF | | LED | | | | | | WS | WD |
| 0.001 | | LL | tFL | | | RF | | LED | | | | | | WS | WD |
| 0.0005 | 0.4275 | LL | tFL | | | RF | | LED | | | | | | WS | WD |
| 0.0001 | | LL | | | | RF | | LED | | | | | | WS | WD |
| 0.00005 | 0.4125 | LL | | | | RF | | LED | | | | | | WS | WD |
| 0.00001 | | LL | | | | | | LED | | | | | | WS | WD |
| 0.000005 | | LL | | | | | | LED | | | | | | WS | WD |
| 0.000001 | 0.3822 | LL | | | | | | LED | | | | | | WS | WD |
| end | 0.3822 | LL | | | | | | LED | | | | | | WS | WD |

**backward elimination steps for pt 'ate_LED_0'**

| limit p-value | R2 | vars included | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 5 | 6-10 | 11 | 13-16 | 17 | 18 | 19 | 20 | 21 | 24 | 33-39 |
| all included | 0.4548 | LL | tFL | tSP | | RF | C | | RE | S | TOD | PM_BG | T | WS | WD |
| 0.05 | 0.4527 | LL | | tSP | | RF | C | | | S | | PM_BG | T | WS | WD |
| 0.01 | | LL | | | | RF | C | | | S | | PM_BG | | WS | WD |
| 0.005 | 0.4512 | LL | | | | RF | C | | | S | | PM_BG | | WS | WD |
| 0.001 | | LL | | | | RF | C | | | S | | PM_BG | | WS | WD |
| 0.0005 | 0.4453 | LL | | | | RF | C | | | S | | PM_BG | | WS | WD |
| 0.0001 | | LL | | | | RF | C | | | S | | PM_BG | | WS | WD |
| 0.00005 | 0.4390 | LL | | | | RF | C | | | S | | PM_BG | | WS | WD |
| 0.00001 | | LL | | | | RF | C | | | S | | | | WS | WD |
| 0.000005 | | LL | | | | RF | C | | | S | | | | WS | WD |
| 0.000001 | 0.4328 | LL | | | | RF | C | | | S | | | | WS | WD |
| end | 0.3973 | LL | | | | RF | C | | | | | | | WS | WD |

**backward elimination steps for pt 'ate_LED_1'**

| limit p-value | R2 | vars included | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 5 | 6-10 | 11 | 13-16 | 17 | 18 | 19 | 20 | 21 | 24 | 33-39 |
| all included | 0.3514 | LL | tFL | tSP | SPL | RF | C | | RE | S | TOD | PM_BG | T | WS | WD |
| 0.05 | | LL | tFL | | SPL | RF | C | | RE | S | | PM_BG | T | WS | WD |
| 0.01 | | LL | tFL | | SPL | RF | C | | RE | S | | PM_BG | T | WS | WD |
| 0.005 | 0.3506 | LL | tFL | | SPL | RF | C | | RE | S | | PM_BG | T | WS | WD |
| 0.001 | | LL | tFL | | | RF | C | | RE | | | | T | WS | WD |
| 0.0005 | 0.3305 | LL | tFL | | | RF | C | | RE | | | | T | WS | WD |
| 0.0001 | | LL | | | | RF | C | | | | | | T | WS | |
| 0.00005 | | LL | | | | RF | C | | | | | | T | WS | |
| 0.00001 | | LL | | | | RF | C | | | | | | T | WS | |
| 0.000005 | | LL | | | | RF | C | | | | | | T | WS | |
| 0.000001 | 0.2937 | LL | | | | RF | C | | | | | | T | WS | |
| end | 0.2937 | LL | | | | RF | C | | | | | | T | WS | |

**Table C-5 Results for steps of backward elimination analysis for grouped ultrafine particle concentration data - ATE model**

**backward elimination steps for pt 'atek_summer'**

| limit p-value | R2 | vars included | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 5 | 6-10 | 11 | 13-16 | 17 | 18 | 19 | 20 | 21 | 24 | 33-39 | |
| all included | 0.3835 | LL | tFL | tSP | SPL | RF | C | LED | RE | | TOD | PM_BG | T | WS | WD | |
| 0.05 | 0.3818 | LL | | tSP | SPL | RF | | LED | | | | PM_BG | T | WS | WD | |
| 0.01 | 0.3785 | LL | | tSP | | RF | | LED | | | | PM_BG | T | WS | WD | |
| 0.005 | 0.3746 | LL | | tSP | | RF | | LED | | | | PM_BG | | WS | WD | |
| 0.001 | | LL | | tSP | | RF | | LED | | | | PM_BG | | WS | | |
| 0.0005 | | LL | | tSP | | RF | | LED | | | | PM_BG | | WS | | |
| 0.0001 | 0.3614 | LL | | tSP | | RF | | LED | | | | PM_BG | | WS | | |
| 0.00005 | | LL | | | | RF | | LED | | | | PM_BG | | WS | | |
| 0.00001 | | LL | | | | RF | | LED | | | | PM_BG | | WS | | |
| 0.000005 | | LL | | | | RF | | LED | | | | PM_BG | | WS | | |
| 0.000001 | 0.3520 | LL | | | | RF | | LED | | | | PM_BG | | WS | | |
| end | 0.3319 | LL | | | | RF | | LED | | | | | | WS | | |

**backward elimination steps for pt 'atek_winter'**

| limit p-value | R2 | vars included | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 5 | 6-10 | 11 | 13-16 | 17 | 18 | 19 | 20 | 21 | 24 | 33-39 | |
| all included | 0.4374 | LL | tFL | tSP | SPL | RF | C | LED | RE | | TOD | PM_BG | T | WS | WD | |
| 0.05 | 0.4297 | LL | tFL | | | RF | | LED | | | | | T | WS | WD | |
| 0.01 | | LL | tFL | | | RF | | LED | | | | | | WS | WD | |
| 0.005 | 0.4235 | LL | tFL | | | RF | | LED | | | | | | WS | WD | |
| 0.001 | | LL | | | | RF | | LED | | | | | | WS | WD | |
| 0.0005 | | LL | | | | RF | | LED | | | | | | WS | WD | |
| 0.0001 | | LL | | | | RF | | LED | | | | | | WS | WD | |
| 0.00005 | 0.4117 | LL | | | | RF | | LED | | | | | | WS | WD | |
| 0.00001 | | LL | | | | | | LED | | | | | | WS | WD | |
| 0.000005 | | LL | | | | | | LED | | | | | | WS | WD | |
| 0.000001 | 0.3781 | LL | | | | | | LED | | | | | | WS | WD | |
| end | 0.3781 | LL | | | | | | LED | | | | | | WS | WD | |

**backward elimination steps for pt 'atek_LED_0'**

| limit p-value | R2 | vars included | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 5 | 6-10 | 11 | 13-16 | 17 | 18 | 19 | 20 | 21 | 24 | 33-39 | |
| all included | 0.4949 | LL | tFL | tSP | | RF | C | | RE | S | TOD | PM_BG | T | WS | WD | |
| 0.05 | 0.4945 | LL | | tSP | | RF | C | | | S | | PM_BG | T | WS | WD | |
| 0.01 | | LL | | | | RF | C | | | S | | PM_BG | | WS | WD | |
| 0.005 | | LL | | | | RF | C | | | S | | PM_BG | | WS | WD | |
| 0.001 | | LL | | | | RF | C | | | S | | PM_BG | | WS | WD | |
| 0.0005 | | LL | | | | RF | C | | | S | | PM_BG | | WS | WD | |
| 0.0001 | | LL | | | | RF | C | | | S | | | | WS | WD | |
| 0.00005 | 0.4880 | LL | | | | RF | C | | | S | | | | WS | WD | |
| 0.00001 | | LL | | | | RF | C | | | S | | | | WS | WD | |
| 0.000005 | 0.4736 | LL | | | | RF | C | | | S | | | | WS | WD | |
| 0.000001 | 0.4526 | LL | | | | RF | | | | S | | | | WS | WD | |
| end | 0.4526 | LL | | | | RF | | | | S | | | | WS | WD | |

**backward elimination steps for pt 'atek_LED_1'**

| limit p-value | R2 | vars included | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 5 | 6-10 | 11 | 13-16 | 17 | 18 | 19 | 20 | 21 | 24 | 33-39 | |
| all included | 0.3425 | LL | tFL | tSP | SPL | RF | C | | RE | S | TOD | PM_BG | T | WS | WD | |
| 0.05 | 0.3404 | LL | tFL | tSP | | RF | C | | RE | S | | PM_BG | T | WS | WD | |
| 0.01 | | LL | | | | RF | C | | | S | | PM_BG | T | WS | WD | |
| 0.005 | 0.3313 | LL | | | | RF | C | | | S | | PM_BG | T | WS | WD | |
| 0.001 | | LL | | | | RF | C | | | | | | T | WS | WD | |
| 0.0005 | 0.3174 | LL | | | | RF | C | | | | | | T | WS | WD | |
| 0.0001 | | LL | | | | RF | C | | | | | | T | WS | | |
| 0.00005 | | LL | | | | RF | C | | | | | | T | WS | | |
| 0.00001 | | LL | | | | RF | C | | | | | | T | WS | | |
| 0.000005 | | LL | | | | RF | C | | | | | | T | WS | | |
| 0.000001 | 0.2964 | LL | | | | RF | C | | | | | | T | WS | | |
| end | 0.2964 | LL | | | | RF | C | | | | | | T | WS | | |

**Table C-6 Results for steps of backward elimination analysis for grouped ultrafine particle concentration data - ATEK model**

**backward elimination steps for pt 'ac_summer'**

| limit p-value | R2 | vars included | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 5 | 6-10 | 11 | 13-16 | 17 | 18 | 19 | 20 | 21 | 24 | 33-39 |
| all included | 0.3685 | LL | tFL | tSP | SPL | RF | C | LED | RE | | TOD | PM_BG | T | WS | WD |
| 0.05 | 0.3665 | | | tSP | | RF | | LED | RE | | TOD | PM_BG | T | WS | WD |
| 0.01 | 0.3575 | | | | | RF | | | RE | | TOD | PM_BG | T | WS | WD |
| 0.005 | 0.3533 | | | | | RF | | | RE | | | PM_BG | T | WS | WD |
| 0.001 | | | | | | RF | | | | | | PM_BG | T | WS | WD |
| 0.0005 | | | | | | RF | | | | | | PM_BG | T | WS | WD |
| 0.0001 | | | | | | RF | | | | | | PM_BG | T | WS | WD |
| 0.00005 | | | | | | RF | | | | | | PM_BG | T | WS | WD |
| 0.00001 | | | | | | RF | | | | | | PM_BG | T | WS | WD |
| 0.000005 | | | | | | RF | | | | | | PM_BG | T | WS | WD |
| 0.000001 | 0.3474 | | | | | RF | | | | | | PM_BG | T | WS | WD |
| end | 0.2993 | | | | | | | | | | | PM_BG | | WS | WD |

**backward elimination steps for pt 'ac_winter'**

| limit p-value | R2 | vars included | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 5 | 6-10 | 11 | 13-16 | 17 | 18 | 19 | 20 | 21 | 24 | 33-39 |
| all included | 0.5600 | LL | tFL | tSP | SPL | RF | C | LED | RE | | TOD | PM_BG | T | WS | WD |
| 0.05 | 0.5534 | | tFL | tSP | SPL | | | LED | | | TOD | PM_BG | T | WS | WD |
| 0.01 | 0.5329 | | tFL | | SPL | | | | | | | PM_BG | T | WS | WD |
| 0.005 | | | | | SPL | | | | | | | | T | WS | WD |
| 0.001 | | | | | SPL | | | | | | | | T | WS | WD |
| 0.0005 | | | | | SPL | | | | | | | | T | WS | WD |
| 0.0001 | 0.5203 | | | | SPL | | | | | | | | T | WS | WD |
| 0.00005 | | | | | | | | | | | | | T | WS | WD |
| 0.00001 | | | | | | | | | | | | | T | WS | WD |
| 0.000005 | | | | | | | | | | | | | T | WS | WD |
| 0.000001 | 0.5051 | | | | | | | | | | | | T | WS | WD |
| end | 0.5051 | | | | | | | | | | | | T | WS | WD |

**backward elimination steps for pt 'ac_LED_0'**

| limit p-value | R2 | vars included | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 5 | 6-10 | 11 | 13-16 | 17 | 18 | 19 | 20 | 21 | 24 | 33-39 |
| all included | 0.4101 | LL | tFL | tSP | | RF | C | | RE | S | TOD | PM_BG | T | WS | WD |
| 0.05 | | | | tSP | | RF | | | | | | PM_BG | T | WS | WD |
| 0.01 | | | | tSP | | RF | | | | | | PM_BG | T | WS | WD |
| 0.005 | 0.4003 | | | tSP | | RF | | | | | | PM_BG | T | WS | WD |
| 0.001 | | | | | | | | | | | | | T | WS | WD |
| 0.0005 | | | | | | | | | | | | | T | WS | WD |
| 0.0001 | | | | | | | | | | | | | T | WS | WD |
| 0.00005 | | | | | | | | | | | | | T | WS | WD |
| 0.00001 | | | | | | | | | | | | | T | WS | WD |
| 0.000005 | | | | | | | | | | | | | T | WS | WD |
| 0.000001 | 0.3714 | | | | | | | | | | | | T | WS | WD |
| end | 0.3714 | | | | | | | | | | | | T | WS | WD |

**backward elimination steps for pt 'ac_LED_1'**

| limit p-value | R2 | vars included | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 5 | 6-10 | 11 | 13-16 | 17 | 18 | 19 | 20 | 21 | 24 | 33-39 |
| all included | 0.4967 | LL | tFL | tSP | SPL | RF | C | | RE | S | TOD | PM_BG | T | WS | WD |
| 0.05 | 0.4955 | | tFL | | SPL | RF | | | | S | TOD | PM_BG | T | WS | WD |
| 0.01 | | | tFL | | | RF | | | | S | TOD | PM_BG | T | WS | WD |
| 0.005 | | | tFL | | | RF | | | | S | TOD | PM_BG | T | WS | WD |
| 0.001 | 0.4925 | | tFL | | | RF | | | | S | TOD | PM_BG | T | WS | WD |
| 0.0005 | 0.4830 | | tFL | | | RF | | | | | | PM_BG | T | WS | WD |
| 0.0001 | | | tFL | | | RF | | | | | | PM_BG | T | WS | WD |
| 0.00005 | | | tFL | | | RF | | | | | | PM_BG | T | WS | WD |
| 0.00001 | | | | | | RF | | | | | | PM_BG | T | WS | WD |
| 0.000005 | | | | | | RF | | | | | | PM_BG | T | WS | WD |
| 0.000001 | 0.4738 | | | | | RF | | | | | | PM_BG | T | WS | WD |
| end | 0.4579 | | | | | RF | | | | | | | T | WS | WD |

**Table C-7 Results for steps of backward elimination analysis for grouped ultrafine particle concentration data - AC model**

Appendix

# D Auxiliary Material for Chapter 6

## D.1 Interpolation of missing values in the field study matrix

In order to estimate annual rush hour exposure, average journey exposure values were compiled in a two-dimensional matrix, using the corresponding temperature and wind speed values to define the bin values in the first and second dimension, respectively. Figures D-1a and b show two representations of the matrix, as number distribution (proportion of values in each bin) and average values (average of all average exposure values in the bin). A number of bins remain empty since no field study measurements match the meteorological conditions represented by these bins, particularly for higher wind speeds. Empty bins reduce the amount of field study and reference data that can be used to estimate annual average exposure as proposed in Chapter 6.2. It was therefore investigated whether standard interpolation methods could be used to 'fill' the empty bins.

Interpolation of missing bin values always introduces inaccuracies and may even skew the data, particularly when whole sections of the matrix consist of empty bins. In order to minimise accuracies, only bins which had values in at least three of the four closest neighbouring bins were considered for interpolation here. Neighbouring bins were defined here as the bins directly above, below, left or right of the empty bin. In this case there are four empty bins with four valid neighbouring bins and four bins with three neighbours. However, the empty bin with three neighbours closest to the x

axis was not included since there were no values in the corresponding bin in the reference matrix, which indicates that those conditions do not typically occur. For the remaining seven bins, it is reasonable to assume that if wind speed and temperature conditions represented by these bins had occurred during the data collection phase, the concentrations would have been similar to the ones recorded for similar conditions, i.e. as shown by the values in the surrounding bins.

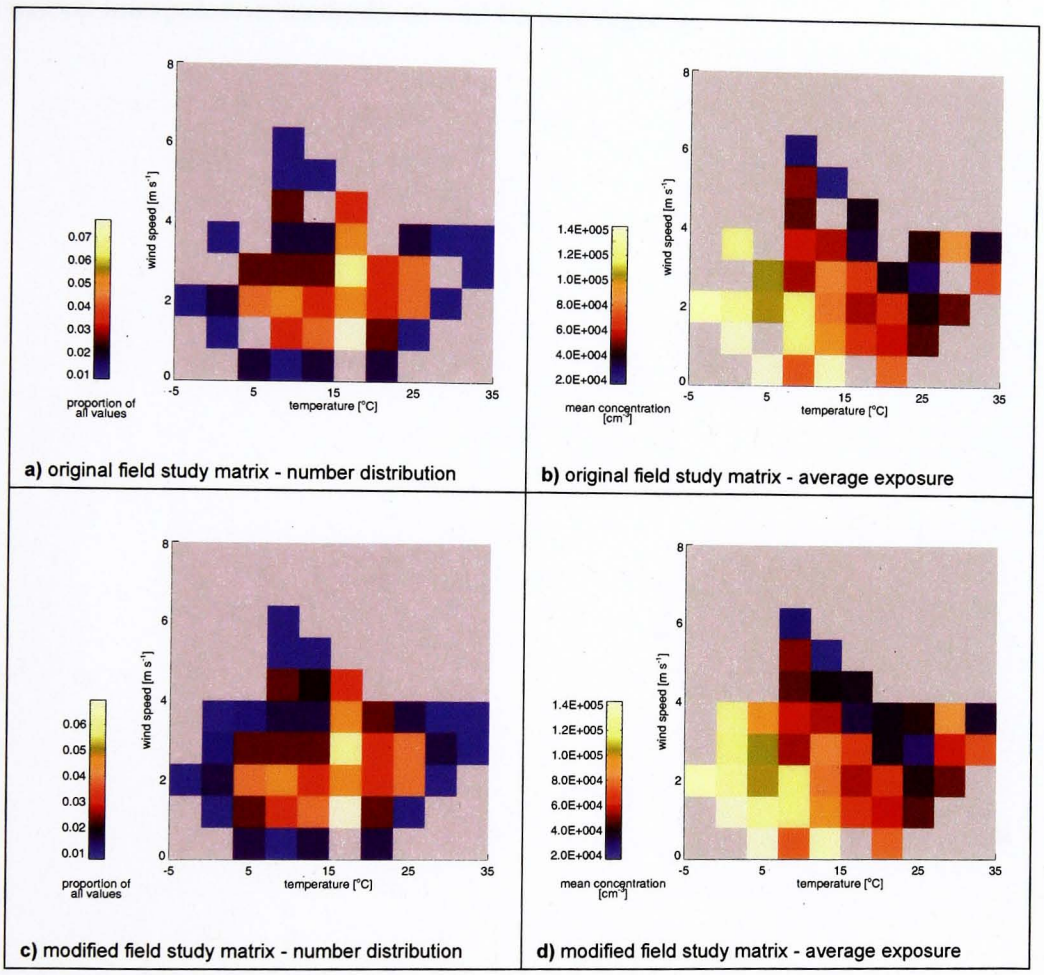


**Figure D-1 Matrices of field study data before and after interpolation**

By calculating the average of the values in the three or four closest neighbouring bins, i.e. disregarding empty bins, a value was generated for each of the seven bins considered for interpolation. A new number distribution matrix was also generated, using the same method but including empty neighbouring bins to interpolate bin values. The resulting matrices are shown in Figures D-1c and d.

By applying this simple interpolation method to only a few bins of the field study matrix, the correlation between number distributions of field study matrix and the meteorological reference matrix could be increased from $r^2 = 0.33$ for the original matrix to $r^2 = 0.45$ for the modified matrix.

Although in this case, the interpolation method improved the correlation between field study and reference data, it may not be suitable for other data sets, particularly if values are to be interpolated for a larger number of empty bins. In such cases, more advanced interpolation methods should be investigated.